JRC Scientific and Technical Reports

Institute for Energy and Transport

ENIQ TGR TECHNICAL DOCUMENT

Influence of Sample Size and Other Factors on Hit/Miss Probability of Detection Curves

ENIQ report No 47

Charles Annis & Luca Gandossi

JRC EUROPEAN COMMISSION

iet

EUR 25200 EN – 2012

The mission of the JRC-IE is to provide support to Community policies related to both nuclear and non-nuclear energy in order to ensure sustainable, secure and efficient energy production, distribution and use.

European Commission
Joint Research Centre
Institute for Energy

Contact information
Address: Westerduinweg 3, NL-1755 LE Petten
E-mail: luca.gandossi@jrc.nl
Tel.: +31.224.565250
Fax: +31.224.565641

http://iet.jrc.ec.europa.eu/
http://ec.europa.eu/dgs/jrc/index.cfm

JRC68677

EUR 25200 EN
ISBN 978-92-79-23018-9
ISSN 1831-9424
doi:10.2790/43050

Luxembourg: Publications Office of the European Union, 2012

European Commission
Directorate General Joint Research Centre
Institute for Energy
Petten, The Netherlands

ENIQ TGR TECHNICAL DOCUMENT

INFLUENCE OF SAMPLE SIZE AND OTHER FACTORS ON HIT/MISS PROBABILITY OF DETECTION CURVES

January 2012 ***ENIQ Report nr. 47*** ***EUR 25200 EN***

ENIQ, the European Network for Inspection and Qualification, publishes three types of documents:

Type 1 — **Consensus documents**
Consensus documents contain harmonised principles, methods, approaches and procedures and emphasize the degree of harmonisation between ENIQ members.

Type 2 — **Position/Discussion documents**
Position/discussion documents contain compilations of ideas, express opinions, review practices, draw conclusions and make recommendations for technical projects.

Type 3 — **Technical reports**
Technical reports contain results of investigations, compilations of data, reviews and procedures without expressing any specific opinion or evaluation on behalf of ENIQ.

This 'ENIQ TGR Technical Document – Influence of Sample Size and Other Factors on Hit/Miss Probability of Detection Curves' (ENIQ Report No 47) is a type 3 document.

FOREWORD

The present work is the outcome of the activities of the ENIQ Task Group on Risk (TGR).

ENIQ, the European Network for Inspection and Qualification, is driven by the nuclear utilities in the European Union and Switzerland and managed by the European Commission's Joint Research Centre (JRC). It is active in the field of in-service inspection (ISI) of nuclear power plants by non-destructive testing (NDT), and works mainly in the areas of qualification of NDT systems and risk-informed in-service inspection (RI-ISI). This technical work is performed in two task groups: TG Qualification and TG Risk.

A key achievement of ENIQ has been the issuing of a European Methodology Document for Inspection Qualification, which has been widely adopted across Europe. This document defines an approach to the qualification of inspection procedures, equipment and personnel based on a combination of technical justification (TJ) and test piece trials (open or blind). The TJ is a crucial element in the ENIQ approach, containing evidence justifying that the proposed inspection will meet its objectives in terms of flaw detection and sizing capability. The assurance provided is nonetheless often qualitative. Obtaining a quantitative measure of inspection reliability is becoming more and more important, as structural reliability modelling and quantitative risk-informed in-service inspection methodologies become more widely used within the nuclear industry in Europe. Such a measure is essential to quantify the reduction of failure probability, and hence risk reduction, after inspection.

The purposes of this document, aimed mostly at NDT engineers and practitioners, are threefold: (1) to extend the conclusions of an earlier report (ENIQ report No 41: "Probability of Detection Curves: Statistical Best-Practices"), (2) to justify the Rule-of-Thumb that a valid Probability of Detection (POD) *vs.* size curve requires a minimum of 60 targets for binary response (hit/miss) data, (3) to provide guidelines for the NDE practitioner in designing a study to assess the effectiveness of a binary response inspection system using POD *vs.* size curves.

The active members of the ENIQ Task Group on Risk are (in alphabetical order):

D. Couplet	Tractebel, Belgium
L. Gandossi	JRC, European Commission, the Netherlands
J. Gunnars	Inspecta Oy, Sweden
L. Horacek	NRI, Czech Republic
E. Kichev	Kozloduy NPP, Bulgaria
P. Lafrenière	CANDU Owners Group, Canada
A. Leijon	Ringhals AB, Sweden
P. Luostarinen	Fortum Engineering Ltd, Finland
T. Meister	Ringhals AB, Sweden
P. O'Regan	EPRI, United States
C. Schneider	The Welding Institute, United Kingdom
K. Simola	VTT, Finland
A. Toft	Serco Assurance, United Kingdom
R. Van Sonsbeek	Applus RTD Group, Netherlands
I. Virkkunen	Trueflaw Oy, Finland
A. Walker	Rolls-Royce, United Kingdom

The authors of this report are Charles Annis (www.statisticalengineering.com/) and Luca Gandossi (IET JRC).

The voting members of the ENIQ Steering Committee are:

T. Dawood	EDF Energy, United Kingdom
P. Dombret	Tractebel, Belgium
E. Martin	EDF, France
K. Hukkanen	Teollisuuden Voima OY, Finland
R. Schwammberger	Kernkraftwerk Leibstadt, Switzerland
B. Neundorf	Vattenfall Europe Nuclear Energy, Germany
J. Neupauer	Slovenské Elektrárne, Slovakia
S. Pérez	Iberdrola, Spain
A. Richnau	Ringhals AB, Sweden
P. Kopcil	Dukovany NPP, Czech Republic
D. Szabó	Paks NPP, Hungary

The European Commission representatives in ENIQ are L. Gandossi and O. Martin.

TABLE OF CONTENTS

This page is intentionally left blank.

1 Introduction

The use of probability of detection curves to quantify NDT reliability is common in the aeronautical industry, but relatively less so in the nuclear industry, at least in European countries. The main reason for this lies in the very nature of the components being inspected. Sample sizes of inspected flaws tend to be much lower, and it is often very difficult to procure or manufacture representative flaws in test pieces in a high enough number to draw statistical conclusions on the reliability of the NDT system being investigated. Similar considerations led to the development of the ENIQ inspection qualification methodology, based on the idea of the Technical Justification, i.e. a document assembling evidence and reasoning providing assurance that the NDT system is indeed capable of finding the flaws which is designed to detect. The ENIQ methodology has become widely used in many European countries, and is gaining appreciation outside Europe as well, but the assurance it provides is usually of qualitative nature. The need to quantify the output of inspection qualification has become more and more important, especially as structural reliability modelling and quantitative risk-informed in-service inspection methodologies become more widely used. To take full credit of the inspections in structural reliability evaluations, a measure of the NDT reliability is necessary. A probability of detection (POD) curve provides such a metric.

In 2011, the Joint Research Centre supported ENIQ by developing a technical report on Probability of Detection Curves, filling part of the need described above (Gandossi and Annis (2010)). That paper reviewed in a structured way the statistical models that have been proposed to quantify inspection reliability. It is now of interest to investigate further the question of the sample size required to determine a reliable POD curve. Manufacturing or procuring cracks that are representative of real defects found in nuclear power plants can be very expensive when not outright impossible. There is therefore a tendency to reduce sample sizes, in turn increasing the uncertainty associated with the resulting POD curve. Not much guidance on appropriate sample sizes can be found in the published literature, where often this kind of statement is given: "*For hit/miss data experience has shown that 60 specimens is often adequate, and using fewer often results in confidence bounds, while valid, that are too broad to be useful ...*". Such a recommendation is based solely on experience and no formal studies have been published to substantiate it.

The aims of the work summarised in this paper were (1) to develop numerical simulations to determine appropriate and effective inspection target sizes, their number, and distribution to produce valid POD *vs.* size curves, and (2) to summarize these findings as guidelines for the NDE practitioner in designing an experiment to assess system inspection effectiveness.

2 Scope of study

This study considered binary responses (*i.e.* hit or miss) to different size cracks (targets), modelled with a Generalized Linear Model using the logit link. These choices are justified by the following considerations.

The main reason to focus this study on binary response data was the consideration that binary response data and continuous response (*â vs. a*) data are very different: their analysis

methods are different (see Gandossi and Annis, 2010) and lessons learned from one seldom apply to the other. Other reasons for focusing on binary response data were:

1. In practice most NDE tests record a hit/miss response (whereas *â vs. a* data typically only arise for particular ultrasonic and electromagnetic methods).

2. Binary data contain less information than *â* data, which is why MIL-HDBK-1823A (2009) recommends at least 40 *â vs. a* specimens and at least 60 hit/miss specimens.

3. Although some *â vs. a* experiments have used fewer than 40 specimens, the penalty is usually wider confidence bounds. With fewer than 60 hit/miss specimens the practitioner can encounter numerous difficulties: bizarre parameter estimates (*e.g.* POD *vs.* size curves that slope downwards), non-convergence of the GLM algorithm, and very broad confidence bounds.

4. The most serious problem commonly encountered with an *â vs. a* experiment is when all the targets are detected. This is usually because the detection threshold is set too low and much of what is "detected" is noise. This problem has nothing to do with number of samples or target size and distribution.

The generalized linear model (GLM) logit link was selected for several reasons:

1. In practice nearly all binary response NDE data are most effectively modelled using the logit link. The logit is the most common link used in other scientific areas such as medicine and pharmacology.

2. The other commonly used symmetric link, the probit, is overly sensitive to unexpected behaviour in POD extremes (POD close to either zero or one) and often results in models that do not describe the data as effectively as does the logit link.

3. Where an asymmetric link might be indicated, it is often because of data that do not meet the requirements for a two-parameter POD *vs.* size model, *viz.* that the POD approach zero on the left and one on the right. The problem of POD “floor” and/or POD “ceiling” is discussed Gandossi and Annis, 2010, section 3.5.2.

3 BACKGROUND

3.1 The POD *vs.* size model

As stated above, this study considered binary responses (*i.e.* hit or miss) to different size cracks (targets), modelled with a Generalized Linear Model using the logit link. The following overview of the model is included for completeness; Gandossi and Annis (2010) offer a much more complete explanation of these concepts.

There are of course other factors, other than size, that influence probability of detection. For example, characteristics of the target, such as orientation, morphology, density, and chemical composition. There are also other influential factors related to the milieu in which the crack, or other target, is located, such as proximity to the surface, acoustic or electrical properties of the medium, and component shape, including radii of curvature. While all of

these can influence POD, the single most influential factor, ceteris paribus, is always target size, which is why size figures most prominently in POD studies like this one. Whatever else is to be considered, analysis of POD models begin with the influence of size, which provides the foundation for further investigation.

Continuous (uncensored) response data can be modelled using the familiar ordinary least-squares (OLS) regression, see Gandossi and Annis (2010), section 3.3. Binary response data can also be described with a regression model that is a generalization of the linear model. For ordinary regression we say that $y=f(X)$. For hit/miss data, we need some function of y that can *link* (through the probability of the outcome) the binary response to the function of x, $g(y)=f(X)$. This generalization is called a Generalized Linear Model (GLM). Obviously, for ordinary regression, $g(y)=y$.

The most useful (and most widely used) link function is the logistic function (also called *logit* or *log-odds*):

$$f(X) = g(y) = \log\left(p/(1-p)\right) \quad \text{(Eq. 1)}$$

The "odds" are defined as the probability of occurrence of a binary outcome divided by the probability of non-occurrence:

$$odds \equiv \frac{p}{1-p} \quad \text{(Eq. 2)}$$

The log of the odds (hence *log-odds*) is the logit:

$$\log(odds) \equiv \log\left(\frac{p}{1-p}\right) \quad \text{(Eq. 3)}$$

The log-odds POD model is thus

$$\log\left(\frac{POD(a)}{1\text{-}POD(a)}\right) = \beta_0 + \beta_1\, a \quad \text{(Eq. 4)}$$

or also, commonly:

$$\log\left(\frac{POD(a)}{1\text{-}POD(a)}\right) = \beta_0 + \beta_1\, \log(a) \quad \text{(Eq. 5)}$$

Whether or not to transform *size* logarithmically depends almost entirely on the data being modelled, so no universal transformation is recommended. For this reason, we use $h(a)$ to mean either a, or $\log(a)$, depending on the data. Solving (Eq. 5) for POD(a) produces:

$$POD(a,...) = f(a,\boldsymbol{\beta}) = \frac{\exp\left(f(a,\boldsymbol{\beta})\right)}{1+\exp\left(f(a,\boldsymbol{\beta})\right)} \text{ where } f(a,\boldsymbol{\beta}) = \beta_0 + \beta_1 \cdot h(a) \quad \text{(Eq. 6)}$$

Where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. The parameters, $(\beta_0, \beta_1)^T$, have no obvious physical interpretation and so it is convenient to re-parameterize as (Eq. 7):

$$POD(a) = f(a, \boldsymbol{\beta}) = \Phi_{link}^{-1}\left(\frac{x-\mu}{\sigma}\right) \quad \text{(Eq. 7)}$$

where for the logit link:

$$\Phi_{link}\left(\frac{x-\mu}{\sigma}\right) = \log\left(\frac{POD(a)}{1-POD(a)}\right) \quad \text{(Eq. 8)}$$

Now:

$$f_1(x \mid \beta_0, \beta_1) = \log\left(\frac{POD(a)}{1\text{-}POD(a)}\right) = f_2(a \mid z), \text{ where } z = \left(\frac{x-\mu}{\sigma}\right) \quad \text{(Eq. 9)}$$

Since $f_1(x \mid \beta_0, \beta_1) = f_2(a \mid z)$ then:

$$\beta_0 + \beta_1 x = \frac{x-\mu}{\sigma} \quad \text{(Eq. 10)}$$

Solving for (μ, σ) in terms of $(\beta_0, \beta_1)^T$, shows that

$$\frac{x-\mu}{\sigma} = \left(\frac{1}{\sigma}\right)x + \left(\frac{-\mu}{\sigma}\right) \quad \text{(Eq. 11)}$$

which means that

$$\beta_1 = \frac{1}{\sigma} \text{ and } \beta_0 = \frac{-\mu}{\sigma} \quad \text{(Eq. 12)}$$

so that

$$\sigma = \frac{1}{\beta_1} \text{ and } \mu = -\beta_0\sigma \quad \text{(Eq. 13)}$$

μ and σ have useful physical interpretations. μ is the size, or log(size), at which POD = 0.5. σ is the inverse of the GLM regression slope. This is illustrated in Figure 1.

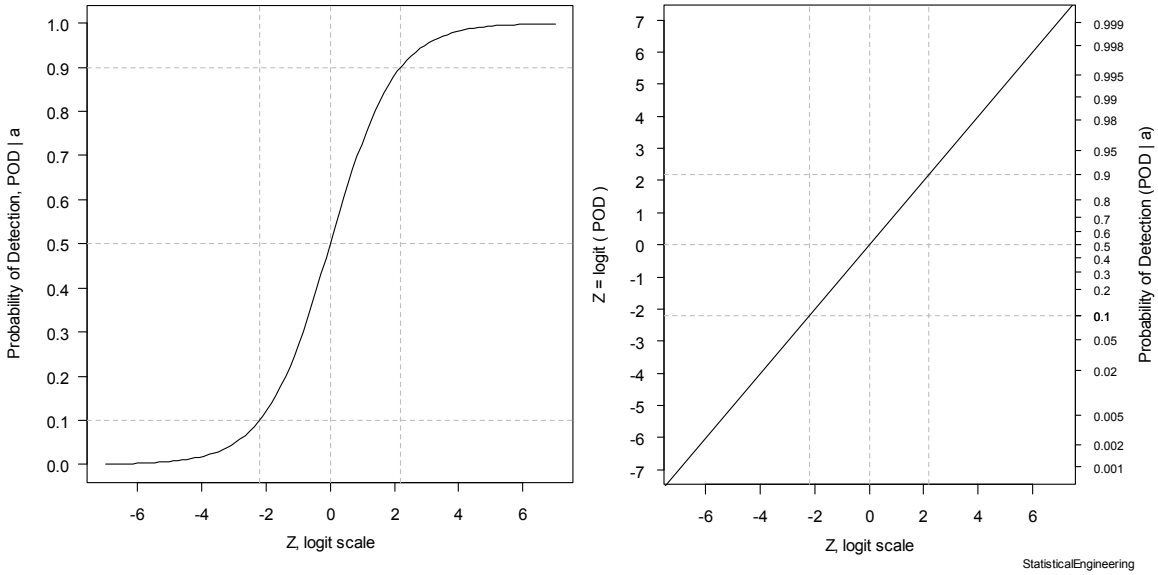


Figure 1
The "S" shaped POD(a) curve plots as a straight line on the logit grid.
The location parameter μ corresponds to the x value at POD=0.5.
The scale parameter σ on the left is 1/slope on the right.

3.2 Coordinate transformation

All regression models having a single explanatory variable can be considered as having zero intercept and unit slope with the data scaled accordingly, but it is standard practice to scale the model parameters, rather than the data. Nonetheless it is helpful to remember that the plot of *Y vs. X* is, essentially zero intercept and unit slope while the axes are scaled to reflect unscaled data.

The logit POD model can also be written, from (Eq. 5), as

$$y = \text{logit}(POD) = f(a, \boldsymbol{\beta}) = \beta_0 + \beta_1 \cdot h(a) \qquad \text{(Eq. 14)}$$

A plot of the logit(POD) *vs.* size for a representative transformation (*e.g.* log), offset and scaling is illustrated in Figure 2.

The scaling in the figures is arbitrary and used only to demonstrate that the underlying generalized linear model is the same regardless of how size is transformed, scaled, or offset. Thus, the results presented herein are applicable without regard to size transformation or scaling. A linear size scale from 0 to 0.4 was used throughout this study for convenience, and it is similar to that in Figure G-42 in MIL-HDBK-1823A.

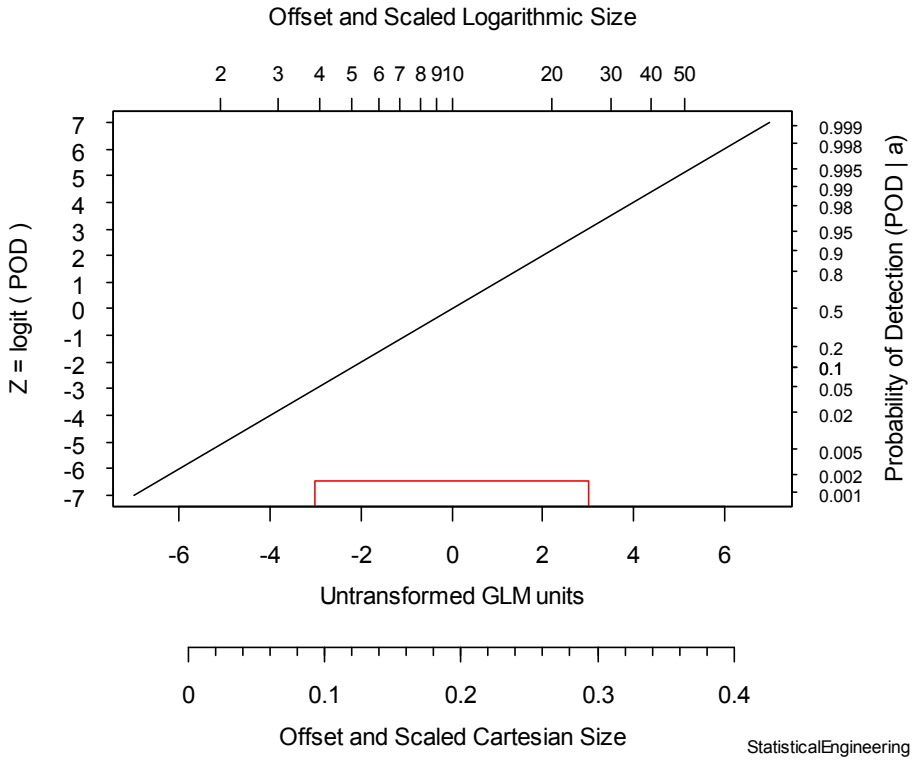


Figure 2
A plot of the logit(POD) is a straight line, as in (Eq. 14)

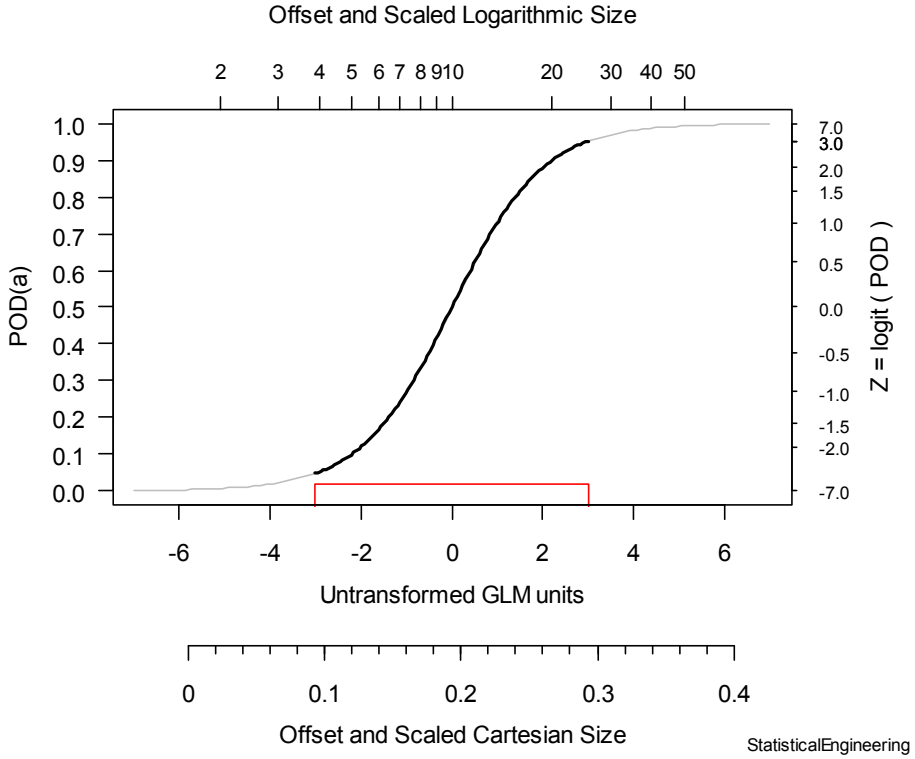


Figure 3
POD plotted vs. offset and scaled size, as in (Eq. 6).

(Eq. 6) gives the probability of detection in explicit terms. This is plotted in Figure 3. The red "box" represents a uniform distribution of sizes. The bold part of the curve is the response in the region for which there is data. The lighter line is an extrapolation.

3.3 Estimation of model parameters and confidence bounds

The model parameters are estimated using the maximum likelihood method, which determines the parameter values that are most likely to have produced the observed experimental results. Parameters estimated in this way are called Maximum Likelihood Estimators (MLEs).

To construct confidence bounds, let us compare MLEs with other possible parameter estimates. If we choose slightly different values, the resulting likelihood gets smaller (in other words, these parameter values are less likely to have produced the observed experimental results). In our case, with two model parameters, the likelihood can be visualised as a surface. In correspondence of the MLEs, the likelihood achieves its maximum. Values of the parameters that are "close" to the best estimates are plausible, and values that are "far" are unlikely to describe the data. The loglikelihood ratio (defined as the ratio of the logarithm of the likelihood evaluated at the "new" values to the logarithm of the maximum value) provides a criterion to express this concept of "closeness", and thus provides a means for constructing likelihood ratio confidence bounds on the POD *vs.* size curve.

It can be shown that, as the sample size increases, the log-likelihood ratio has an asymptotic chi-square distribution, with degrees-of-freedom equal to the number of parameters in the model, specifically:

$$-2\log\left(\frac{L(\theta_0)}{L(\theta)}\right) \sim \chi^2_{1-\alpha;df} \qquad \text{(Eq. 15)}$$

The POD *vs.* size parameters, β_0, β_1, in Figure 4 are the maximum likelihood values at the centre of Figures 5 and 6, indicated with a red plus (+) symbol. Any parameter pair on the loglikelihood contour would produce a single POD *vs.* size curve in Figure 4. Using points along the 95% confidence contour produces many POD *vs.* size curves and the dotted lines in Figure 4 enclose them. Thus we have constructed two-sided 95% confidence bounds on the POD *vs.* size curve by insisting that parameter pairs in Figures 5 and 6 be no further from their maximum likelihood values than specified by the likelihood ratio criterion[1].

An animated demonstration of the relationship between the contours of the loglikelihood surface and the resulting confidence bounds on the POD *vs.* size curve can be seen on-line at http://StatisticalEngineering.com/mh1823/QNDE/mh1823-confidence.html

[1] It has been suggested that the 95% bound should more properly be labelled "97½%" because for very large numbers of samples (N>10,000) the coverage approaches an asymptotic value of 97½%. Changing the label would not change the fact that for reasonable numbers of samples the loglikelihood ratio confidence construction produces a value a 95% confidence value for a_{90} (called $a_{90/95}$) that will be greater than the true a_{90} in 95% of similar experiments. Trying to explain why we used a "97½%" label for the bound that produces 95% coverage for reasonable numbers of samples would, we believe, only confuse the reader. Therefore we will observe the conventional, and accurate, nomenclature of 95%.

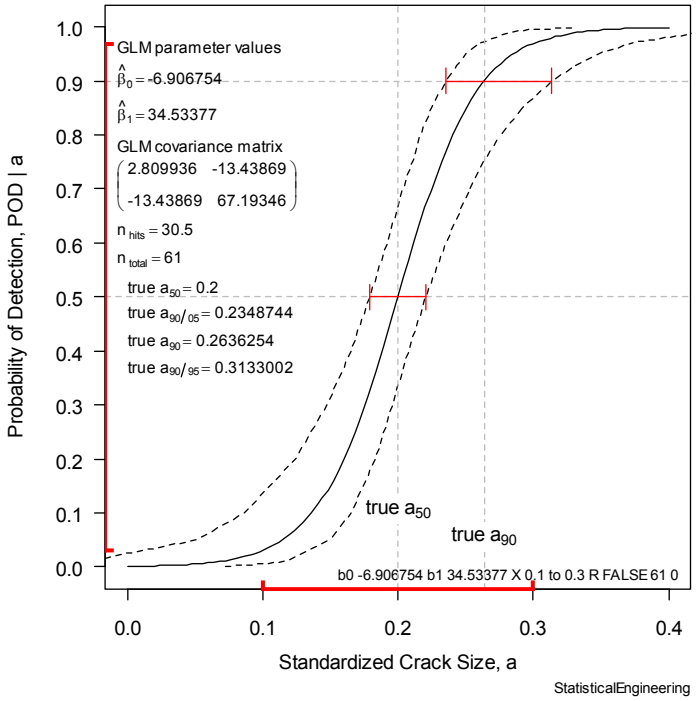


Figure 4
Representative POD vs. Size Curve

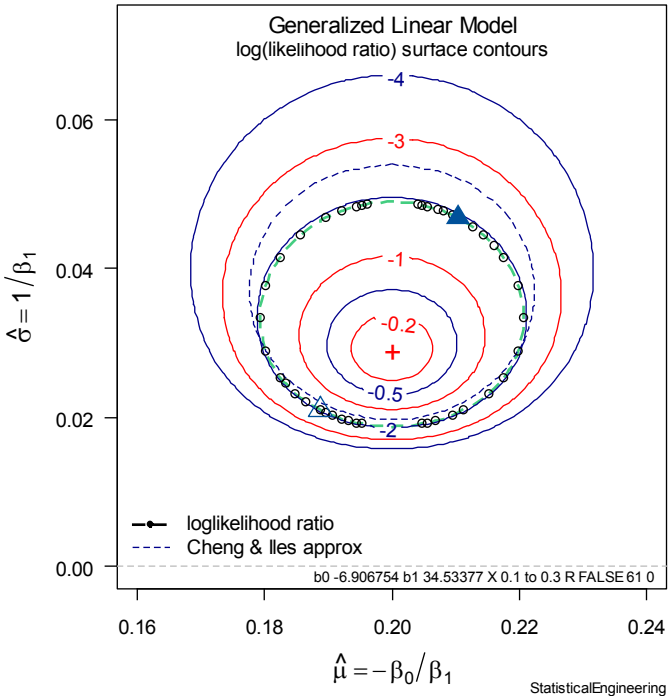


Figure 5
Loglikelihood surface, μ, σ parameterization, showing 95% confidence contour

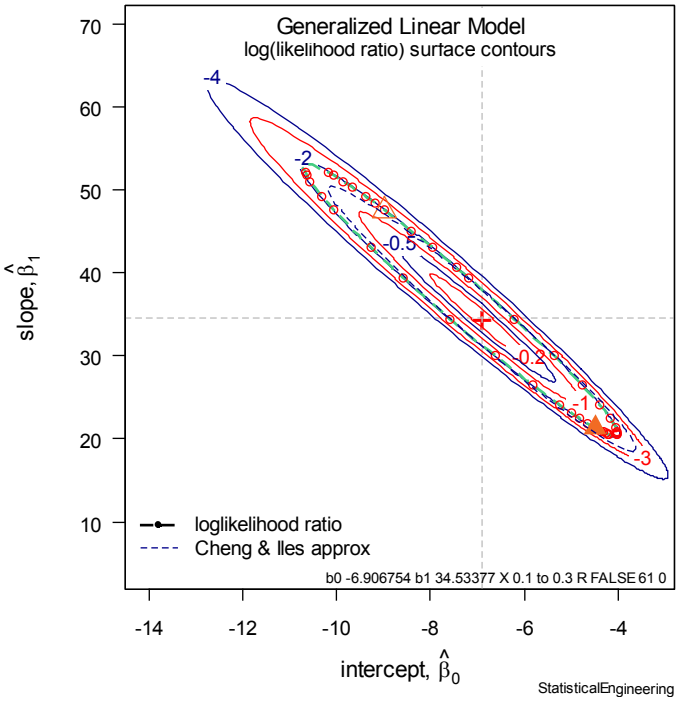


Figure 6
Loglikelihood surface, β_0, β_1 parameterization.

4 METHODOLOGY

4.1 Monte Carlo method

Our original approach was intended to be a straightforward Monte Carlo study to determine the effects of parameters of interest by repeated random sampling and building logistic POD *vs.* size models of each sampling. The following methodology was devised:

1. A true underlying POD *vs.* size relationship is postulated, assuming the model described in Section 3.1, (Eq. 5), and assigning some given values to the (β_0, β_1) pair, thus fully describing the POD curve.

2. Fix a sample size (for instance *N*=30, 45, 60, etc.) and choose a set of crack sizes with some specified distribution.

3. From the postulated POD *vs.* size relationship, and for the set of crack sizes specified above, generate a high number, *m*, of realizations of random hit/miss detection data.

4. For each realization, find the best parameter estimates (*i.e.*, the Maximum Likelihood Estimators, $\hat{\beta}_0, \hat{\beta}_1$).

5. From the spread of estimates ($\hat{\beta}_0, \hat{\beta}_1$) around the true value (β_0, β_1) and from the characteristics of the estimated POD curves when compared with the true POD curve, draw conclusions and give guidance on the sample size required to achieve adequate confidence bounds.

The sample size, *N*, is the number of defects constituting a single NDE experiment. In a real situation, the NDE practitioner manufactures specimens with *N* known defects. The NDE practitioner inspects the *N* defects and records the outcome in binary form. In this work, based on numerical simulation, the Monte Carlo approach is used to simulate a very high number of such experiments for a given sample size. We indicate with *m* the number of simulated experiments. For example, if *N*=30 and we choose *m*=10,000, we simulate 10,000 instances of experiments, each made up of *N*=30 hit/miss data generated from the true POD curve, (Eq. 5).

We are interested in the random, hit/miss response to given cracks (or “targets”). If the cracks change with each Monte Carlo realization, then the influence of the random response is confounded with the influence of not responding to the same set of cracks. It is well known in the statistics community (and less well appreciated among engineers) that random samples from a known probability density do not have all the characteristics of the parent density. An entire area of statistical study deals with "sampling distributions," *i.e.* the behaviour of samples taken from known probability distributions. Perhaps the most familiar is that samples taken from a normal distribution are not themselves normally distributed, but rather follow a Student's "*t*" distribution. For sample sizes more than 30 the differences are not usually important but for 30 or fewer the differences can be dramatic. We decided to avoid this difficulty altogether and thus we elected to consider only the influences of the binary response to known stimuli, that is, responses to the same set of targets.

We initially set about to investigate the number of simulations, *m*, required for the approach described above. This is explained in section 4.5. This early investigation revealed a fundamental difficulty of Monte Carlo (MC) simulation. Because MC studies are random, there remains a random component in their results. If the influence of randomness is on the same order as the influence of the phenomenon under study then enormous numbers of simulations are required to distinguish what is interesting from what is only random. This issue is investigated in sections 4.6 and 4.7. Thus, a new method was devised, as explained in Section 4.8.

4.2 Factors influencing the POD *vs.* size relationship

For any real inspection, the true relationship of probability of detection with target size is unknown and the factors of the NDE experiment have no influence at all on that relationship. The purpose of the experiment is to provide a credible model of POD *vs.* size, and appropriate choices of experimental factors determine how effective that model is.

Although not the subject of this report, it is worth reminding the reader that, in general, POD is influenced by more than one variable, and that the text matrix needs to reflect this, at least by ensuring a balanced design with respect to any ‘nuisance variables’ (ref. MIL-HDBK-1823 Appendix E). This is especially important for Ultrasonic Testing of welding flaws in thick section nuclear components where the orientation (tilt and skew) of planar flaws is known to be highly influential (see ENIQ Recommended Practice 1 ‘Influential/essential parameters’). It should also be remembered that “size” can be more involved that one linear dimension. It can be, for example, the square root of the cross-sectional area normal to the interrogating UT beam.

Our study considered the following influences on the mathematical description of the POD *vs.* size relationship:

1. The number of targets. We considered 30, 45, 60, 90, 120, and 500 (for some studies as many as 5,000).

2. The target size distribution (uniform, symmetrical, skewed left, skewed right).

3. The location and shape ("intercept" and "slope") of the true POD *vs.* size curve.

4. The location of targets with respect to the (unknown) true POD *vs.* size relationship, e.g. POD "coverage" that results from a given target size distribution. (A large number of targets, almost all of which are found, or missed, are less effective in defining the POD *vs.* size relationship than fewer samples more appropriated located.)

4.3 Comparison criteria

To compare these influences we concentrated on the width of the two most important parts of the POD *vs.* size curve: (1) the size range for 95% confidence at 50% POD and (2) the size range for 95% confidence at 90% POD, more commonly referred to as the width at a_{50} and at a_{90}, viz: $a_{50/95}$ and $a_{90/95}$, respectively[2].

4.4 Effects of POD location, shape and transformations of the size metric

Our study of the influence of location and shape ("intercept" and "slope") of the true POD *vs.* size curve, as well as the influence of transformations of the size metric, e.g. X = log(size), was greatly simplified after realizing that these are only superficial differences - the underlying model in every case is (Eq. 6). The scaling and transformations have no effect on the result, i.e. the width of the confidence intervals. They will have the units of the transformed size of course, but the relative influence of everything else (such as the effect of sample size) is unchanged.

To proceed we at first selected parameter values of β_0 and β_1 and size range (before transformation) from -3 to +3. On the face of it, negative values for size might seem unrealistic, although as logarithms of size negative values are commonplace. Since these choices are entirely arbitrary we chose a size range (after transformation) of 0.2 to 0.4 (undefined units, perhaps inches, for example) because it is similar to the real example G-42 presented in MIL-HDBK-1823A. For all our studies, we selected the following values of β_0 and β_1 *:*

$$\begin{aligned} \beta_0 &= -6.906754 \\ \beta_1 &= 34.53377 \end{aligned} \qquad \text{(Eq. 16)}$$

so that the result would be to plot POD from 0.001 to 0.999. That means the slope and intercept are only to transform some size range of interest into values that will (almost) completely cover POD (it is not feasible to cover POD outside that range since it would require an infinite range of sizes). Moreover POD predictions outside this range are not generally of any practical interest. Even if an NDT response is clearly observable, in practice the possibility of human error sets an upper limit on the POD; Marshall (1982) assumed that this upper limit was 99.5% (for the validated ultrasonic inspection of a nuclear pressure

[2] See note 1.

vessel), whereas Bullough *et al* (2007) cites an upper limit of 99.9% (for radiographic inspection).

The values for slope and intercept corresponding to β_0 and β_1 of (Eq. 16) are:

$$\begin{aligned} \mu &= 0.2 \\ \sigma &= 0.028957 \end{aligned} \qquad \text{(Eq. 17)}$$

4.5 Number of required Monte Carlo simulations for convergence

One objective of this study was to determine the confidence coverage based on the number of specimens and their size distribution. The original plan (using conventional Monte Carlo simulation) required knowing how many simulations are sufficient so that any conclusions drawn are valid and not an artefact of a poorly executed Monte Carlo simulation. To this end, the relationship between sample size, *N*, and number of simulations, *m*, (for that sample size) that converges to the theoretical confidence coverage (*i.e.* 95%) had to be established.

The theoretical behaviour holds when the number of cracks is large. To find out how large this number must be, we conducted 10,000 true Monte Carlo simulations and counted how many times the true $a_{90/95}$ is outside the theoretical 95% confidence bound for different numbers of cracks in the sample.

It is important to note that our initial MC studies used odd numbers of targets (i.e. *N*=61 rather than *N*=60, etc.) because even numbers necessarily omit the centremost size for uniform or symmetric size distributions. (On average, the centremost observation has the largest contribution to the likelihood function, Figure 24.) This turned out to be of miniscule significance for sample sizes of 60 or more and the differences are largely due to having more information in a sample of 61 than in a sample of 60. This comparison is illustrated in Appendix 1. For cosmetic reasons, all later studies used even numbers of targets.

Table 1 summarizes the results of this investigation. Salient features of Table 1 are discussed in the following, and are summarized in Figure 7, which plots the fractions of sample a_{90} values less than the theoretical value for $a_{90/05}$ and sample a_{90} values greater than the theoretical value for $a_{90/95}$.

Table 1
Summary of Monte Carlo simulations evaluating effects of sample size

1	2	3	4	5	6	7	8	9	10	11
N	a_{50}	$a_{90/05}$	a_{90}	$a_{90/95}$	$a_{50}>\mu^{(1)}$	$a_{90}>a_{90}^{(2)}$	$>a_{90/95}^{(3)}$	$>a_{90/95}^{(4)}$	$<a_{90/05}^{(5)}$	$a_{90/95}>a_{90}^{(6)}$
31	0.2	0.2258	0.2636	0.3480	0.5139	0.4170	0.3385	0.0078	0.0752	0.9534
61	0.2	0.2349	0.2636	0.3133	0.4990	0.4540	0.4460	0.0127	0.0581	0.9618
61	0.2	0.2349	0.2636	0.3133	0.5040	0.4498	0.4461	0.0094	0.0580	0.9630
61	0.2	0.2349	0.2636	0.3133	0.5032	0.4522	0.4489	0.0114	0.0550	0.9651
61	0.2	0.2349	0.2636	0.3133	0.5023	0.4538	0.4475	0.0120	0.0582	0.9596
61	0.2	0.2349	0.2636	0.3133	0.4899	0.4484	0.4466	0.0102	0.0611	0.9595
61	0.2	0.2349	0.2636	0.3133	0.5083	0.4541	0.4469	0.0107	0.0546	0.9606
61	0.2	0.2349	0.2636	0.3133	0.4899	0.4484	0.4466	0.0102	0.0611	0.9595
61	0.2	0.2349	0.2636	0.3133	0.5051	0.4531	0.4468	0.0106	0.0555	0.9599
61	0.2	0.2349	0.2636	0.3133	0.4949	0.4417	0.4395	0.0112	0.0587	0.9599
121	0.2	0.2421	0.2636	0.2951	0.4999	0.4629	0.4579	0.0141	0.0466	0.9676
501	0.2	0.2523	0.2636	0.2773	0.4989	0.4902	0.4839	0.0175	0.0320	0.9743
1001	0.2	0.2554	0.2636	0.2730	0.4919	0.4874	0.4840	0.0203	0.0325	0.9712
1001	0.2	0.2554	0.2636	0.2730	0.4994	0.4913	0.4866	0.0197	0.0312	0.9715
2001	0.2	0.2578	0.2636	0.2701	0.5021	0.4946	0.4759	0.0201	0.0298	0.9721
2001	0.2	0.2578	0.2636	0.2701	0.5007	0.4919	0.5258	0.0293	0.0433	0.9718
5001	0.2	0.2600	0.2636	0.2676	0.4976	0.4929	0.4137	0.0216	0.0330	0.9536
5001	0.2	0.2600	0.2636	0.2676	0.5050	0.4984	0.4185	0.0250	0.0364	0.9525

NOTES:
(1) observed fraction of sample values of a_{50} greater than μ_{mle}
(2) observed fraction of sample values of a_{90} greater than true a_{90}
(3) observed fraction of sample values of $a_{90/95}$ greater than true $a_{90/95}$
(4) observed fraction of sample values of a_{90} greater than true $a_{90/95}$
(5) observed fraction of sample values of a_{90} less than true $a_{90/05}$
(6) observed fraction of sample values of $a_{90/95}$ greater than true a_{90}

Column 1 shows the number of targets in the sample, ranging from 31 to 5,001. Samples larger than 200 are of course rare in practice, and were studied here to determine how large a sample was required to achieve asymptotic results. Column 2 shows the true a_{50}, which is 0.2 in our studies. Column 4 reports the target size with 90% POD, a_{90}, which is a particularly interesting quantity. The true lower bound, labelled $a_{90/05}$ is reported in column 3 and the true upper bound, $a_{90/95}$, is shown in column 5.

Note that these labels, $a_{90/05}$ and $a_{90/95}$ are actually misnomers, because the asymptotic bounds are double sided which means that 95% of the observations are expected to fall within them for sufficiently large samples. There are two ways of considering the upper confidence bound, $a_{90/95}$: the first is as the known true value, against which we can compare the 10,000 sample values for a_{90} resulting from the simulations, and the second is as the individually computed $a_{90/95}$ values for each of the 10,000 simulations, to be compared with the true value for a_{90}. The first interpretation is interesting from an asymptotic perspective and, as Table 1 shows, even N=5001 is not large enough to exclude the asymptotic 5% of a_{90} values. On the other hand, the NDE practitioner is less interested in the mathematical fine

points than in the practical performance of the method: we compute a size, which is a sample value of $a_{90/95}$, that is supposed to be larger than the true a_{90}, in about 95 out of 100 similar situations. How well does the quantity perform under this second interpretation? Column 11 shows the fraction of observed $a_{90/95}$ calculations that are greater than a_{90}. The fraction is about 96% (*i.e.* is slightly conservative) and is comparatively insensitive to sample size.

Column 6 shows the fraction of observed a_{50} values that are greater than the true a_{50}. It would be expected that about half of the observed a_{50} values falls on either side of 0.2 and that is what column 6 confirms (Column 6 is a kind of sanity check to show that the simulations are working as they should).

Column 7 is the observed fraction of a_{90} greater than true a_{90}. Again, as with a_{50}, it might appear reasonable to expect that half of the values falls on either side, but the fraction is closer to 0.45 than 0.5 for smaller sample sizes and only approaches symmetry (0.50) for very large samples. Column 8 shows the observed fraction of $a_{90/95}$ greater than true $a_{90/95}$ and also demonstrates a skewed behaviour. This behaviour can be explained by examining Figure 4. The confidence bounds, shown in the figure with dashed lines, are nearly symmetric with respect to the median POD vs. size curve at POD = 50% but are clearly asymmetric at POD = 90%

Columns 9 and 10 consider the left and right confidence bounds on the true a_{90}. For very large samples (N=5,001) these values, summed together, approach 0.05 (5%). Recall that these asymptotic bounds are 95% inclusive.

The goal was to compare the relative sizes of the confidence widths to determine reasonable sample sizes. Achieving the asymptotic value is of less interest, as it would require far more specimens than would ever be practical.

We considered the variability in response for repeated simulations, as shown in Figure 7. This emphasized the lack of precision of true MC sampling. Even repeated simulations for N=5,001 showed noticeable variability in nominally identical simulations, which threatens to mask any underlying trends versus the factors of interest listed in Section 4.2. Fortunately, much of this discussion has only a purely academic interest. The critical fact, demonstrated by these tens of thousands of MC realizations, is that for the practitioner the 95% confidence bounds behave as advertised.

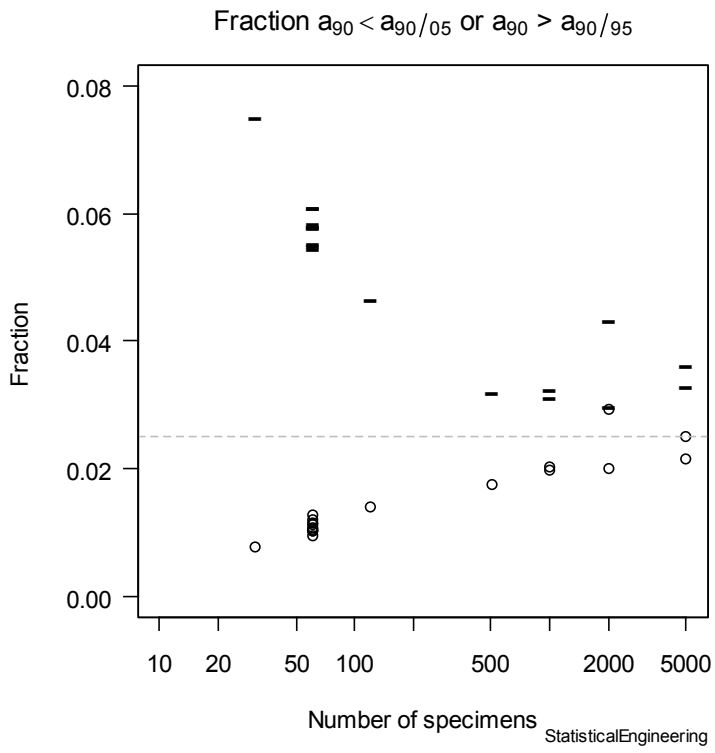


Figure 7
Fraction of a_{90} values less than $a_{90/05}$ or greater than $a_{90/95}$ illustrating the variability due to using "only" 10,000 MC simulations. The asymptotic value of 0.025 apparently requires 5,000 or more cracks,

4.6 Conventional Monte Carlo is too inefficient

The outcome of our early Monte Carlo studies, as shown above (4.5) showed a fundamental difficulty of MC simulation in studying smaller effects, like the distribution shape for a given number of targets, *N*. Because MC studies are random, there remains a random component in their results. If the influence of randomness is on the same order as the influence of the phenomenon under study then enormous numbers of simulations are required to distinguish what is interesting from what is only random and thus not meaningful.

For example, performing a second, identical, study of 10,000, but using a different random number seed, produced a different outcome. Increasing the number of MC realizations might be effective in diminishing the influence of starting random number seed, but the required number of realizations becomes prohibitively large, especially for reasonable numbers of samples (*e.g.* 61 or 121). Even *N*=5,000 samples (a ridiculously large number of crack sizes) may not be sufficient for acceptable precision. To give an example, simulating *m*=10,000 instances of *N*=5,000 sample crack sizes required 14 hours and 16 minutes, and *m*=10,000 may not be an adequate number of simulations.

Nonetheless, in the next section we considered the effect of number of samples on confidence bound widths, based on 10,000 conventional MC realizations.

4.7 Effect of sample size on confidence bound width

The following plots illustrate the effect of sample size (N = 31, 61, 121, 501, 1001, 2001, 5001) on the POD *vs.* size curve, for an increasingly larger sample size. Except for Figure 8, all are based on 10,000 MC realizations. The red box in these plots illustrates the distribution of flaw sizes in the POD experiments (a uniform distribution between a = 0.1 and 0.3, throughout this section). The bold, solid red line indicates the range of flaw sizes (or 'coverage') included in the POD experiments (see also section 4.8.2).

This study illustrates why conventional MC simulation is not feasible. First of all, the total number of computer simulations would require inordinate computer run time. In our investigation, it took about three weeks to produce the plots illustrated in Figures 8–14. These were studies of the effects of sample size, also intended to observe the residual randomness that accompanies MC simulations. Each plot (except Figure 8, which was even more problematic, as explained below) required 10,000 runs and constant checking at the computer as the simulations were accumulated. All this computational effort managed to answer only one of the questions we set out to investigate, *i.e.* the effect on the relative widths of confidence intervals as a function of sample size.

Perhaps more importantly, the large fraction of aberrant datasets that arose at smaller sample sizes (*i.e.* 31 and 61) made it exceedingly tedious to assemble more than a small number of Monte Carlo simulations. An aberrant case is the result of hit/miss data that do not represent the underlying true model. Despite being generated from an assumed true POD curve, an aberrant dataset can occur, not infrequently, purely because of chance. Such situations happen even in practice, which is another argument against using small samples of hit/miss data. The case of N=31 nicely illustrates this point. 10,000 runs simply could not be achieved. Ten MC simulations, starting from different seeds, were required just to accumulate the m=2446 realizations in Figure 8. For reference, the run lengths prior to failure due to aberrant data were: m=108, 2, 12, 781, 229, 256, 327, 302, and 429. Thus, the proportion of simulations that resulted in aberrant data was 0.4%. Similar, albeit less severe, difficulties were encountered for N=61.

Numerical instabilities arose in only 0.4% of the experiments simulated in this report for N=31, but these examples were all for size distributions that were centred on the true a_{50}. Experience suggests that numerical instabilities arise much more often, in practice, where the size distribution is unlikely to be perfectly centred on the true a_{50}.

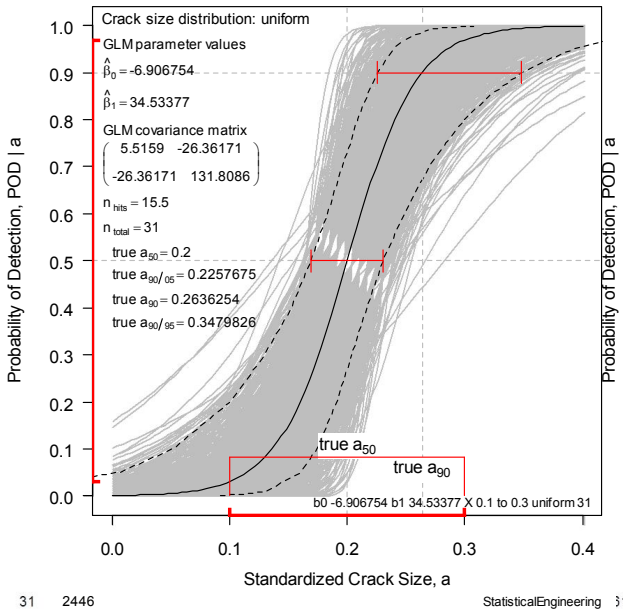


Figure 8
2,446 MC simulations of hit/miss responses to 31 uniformly, evenly distributed crack sizes.

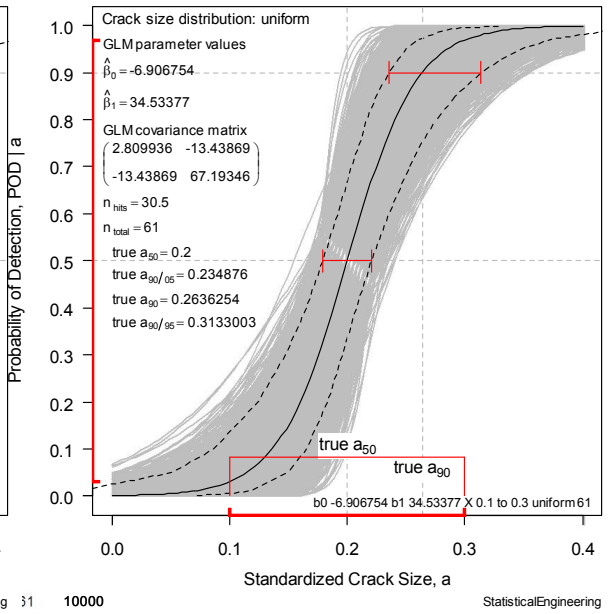


Figure 9
10,000 MC simulations of hit/miss responses to 61 uniformly, evenly distributed crack sizes.

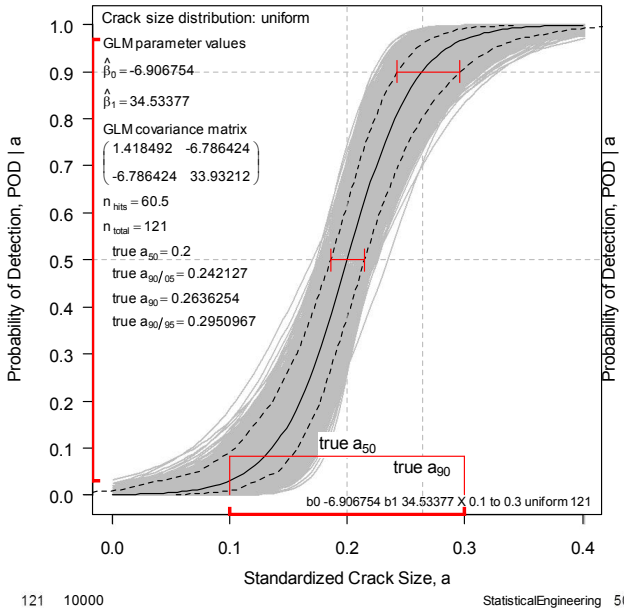


Figure 10
10,000 MC simulations of hit/miss responses to 121 uniformly, evenly distributed crack sizes.

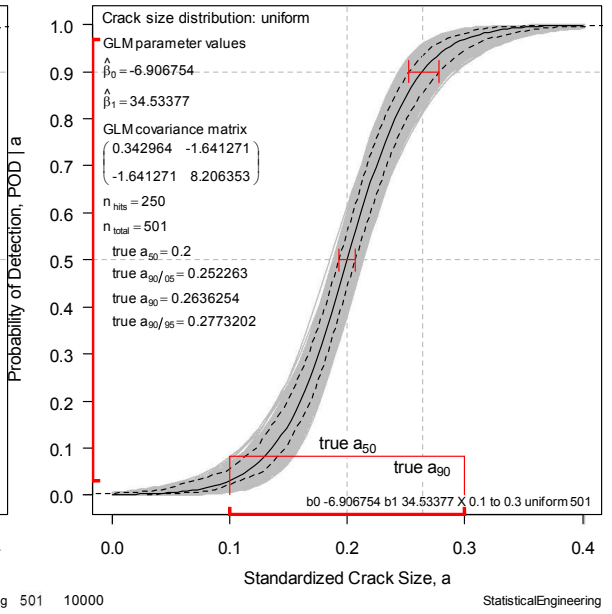


Figure 11
10,000 MC simulations of hit/miss responses to 501 uniformly, evenly distributed crack sizes.

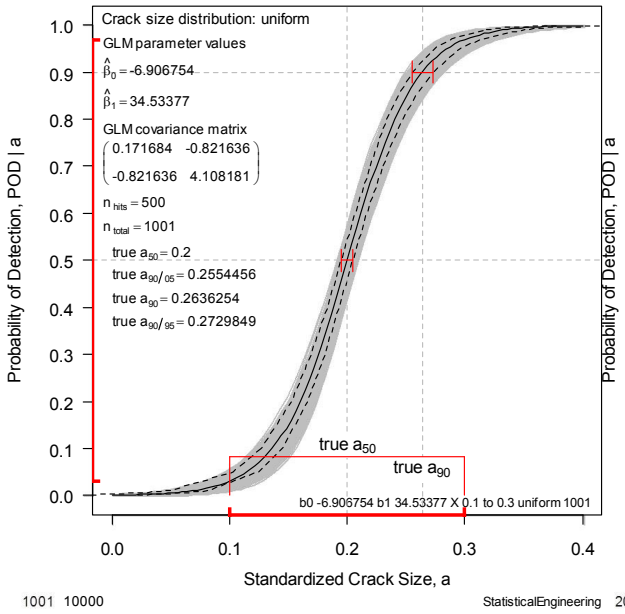


Figure 12
10,000 MC simulations of hit/miss responses to 1001 uniformly, evenly distributed crack sizes

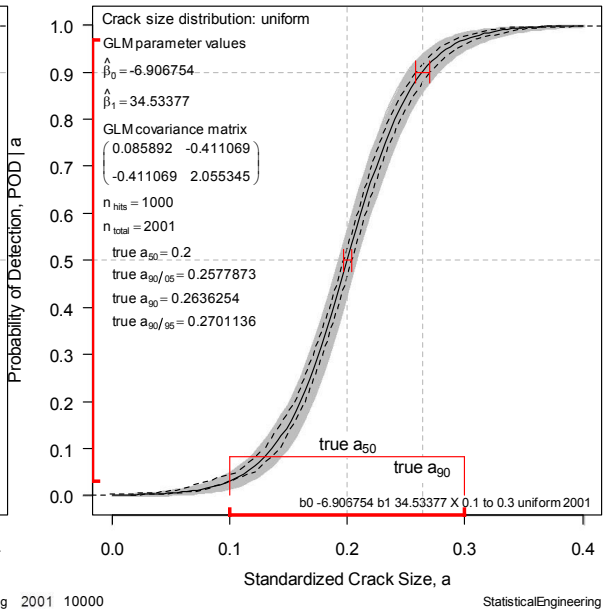


Figure 13
10,000 MC simulations of hit/miss responses to 2001 uniformly, evenly distributed crack sizes

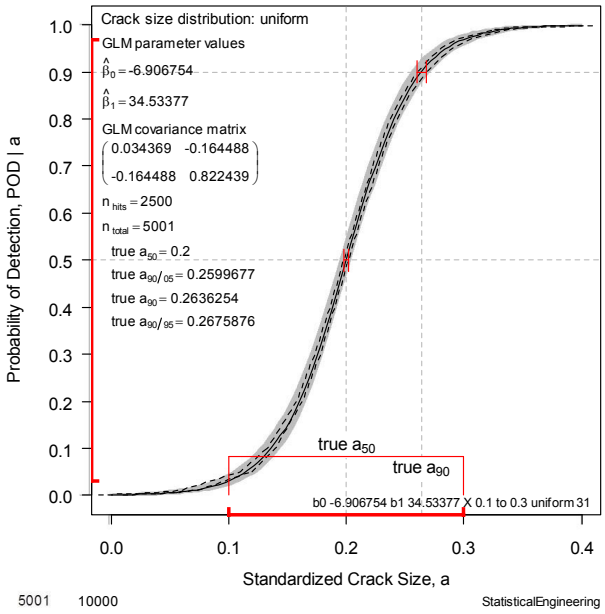


Figure 14
10,000 MC simulations of hit/miss responses to 5001 uniformly, evenly distributed crack sizes

The important features of Figures 8–14 are summarized in Figure 15 by plotting the width of the confidence interval at POD=0.9 as a function of number of cracks in the sample.

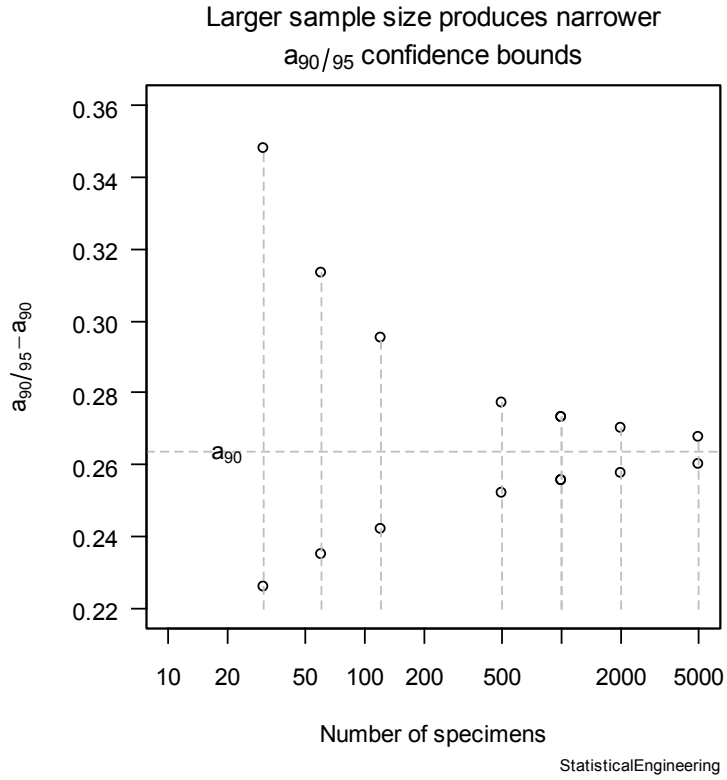


Figure 15
Width of the 95% confidence bound on a_{90} narrows with increasing sample size

Since conventional MC is exceedingly time consuming and encumbered with poor resolution due to inherent randomness, another approach was necessary.

4.8 Devising a surrogate Monte Carlo method

We asked ourselves: why do we need to produce 10,000 realizations? Why not look at the "confidence bounds" on the true model? Because for the true model there is no uncertainty (we know everything about it) so the confidence bounds, even if they existed, would be infinitely narrow. But if we could do such a thing it would eliminate the requirement for all those MC simulations.

Fortunately, we have a more effective approach that not only does away with the excessive computer simulations, but also provides asymptotic (*i.e.* very large sample) results without using very large samples, other than to demonstrate that the approach works.

How can we produce the asymptotic results of, perhaps, millions of Monte Carlo realizations without having to do any MC sampling at all? The new idea is simple: while it is true that confidence bounds on "truth" do not exist (or would be infinitely narrow), we can however consider an idealized, weighted, sample from the true parent POD *vs.* size behaviour. Rather than repeatedly generating random hit/miss responses to a given target size, we will take both hit and miss, and weight them according to their prior likelihood. This requires only a straightforward re-definition of the likelihood function.

It is convenient to describe likelihood as the "probability of the data." This avoids all the statistical hair-splitting that accompanies most textbook definitions, but that definition is not quite correct. It would be more correct to say that "likelihood is the probability that the experiment turned out the way that it did." The reason for this subtle distinction is that likelihood really is the probability of the parameter values, given the data.

Statisticians define *Probability* (of an experimental outcome) as a function of the model parameters. They similarly define *Likelihood* as the probability (of the parameter values) as a function of the data. The mathematical definition (functional form) is the same. The only difference is what is known: the parameter values or the data. If the parameter values are known, and we want to know the outcome of the next experiment, the function defines probability. If the data are known, the function defines the likelihood that the parameter values describe the known data.

The *probability* function for the binomial response model is

$$POD(Y = hit \mid a, \boldsymbol{\beta}) = f(a, \boldsymbol{\beta}) = \frac{\exp\left(f(a, \boldsymbol{\beta})\right)}{1 + \exp\left(f(a, \boldsymbol{\beta})\right)} \text{ where } f(a, \boldsymbol{\beta}) = \beta_0 + a\,\beta_1 \qquad \text{(Eq. 18)}$$

The *likelihood* function for the binomial response model and having the response = hit is

$$L_i(\boldsymbol{\beta} \mid a_i, Y_i = \text{hit}) = f(a_i, \boldsymbol{\beta}) = \frac{\exp\left(f(a_i, \boldsymbol{\beta})\right)}{1 + \exp\left(f(a_i, \boldsymbol{\beta})\right)} \qquad \text{(Eq. 19)}$$

The *likelihood* function for the binomial response model for response = miss is

$$L_i(\boldsymbol{\beta} \mid a_i, Y_i = \text{miss}) = 1 - f(a_i, \boldsymbol{\beta}) = 1 - \frac{\exp\left(f(a_i, \boldsymbol{\beta})\right)}{1 + \exp\left(f(a_i, \boldsymbol{\beta})\right)} \qquad \text{(Eq. 20)}$$

In other words, if the response is *Y*=1 (hit) $L_i = f(a_i, \boldsymbol{\beta})$; if the response is a miss (0), $L_i = 1 - f(a_i, \boldsymbol{\beta})$.

4.8.1 Real Experiments *vs.* Simulations

The foregoing refers to real (not simulated) data, where we know the response, hit or miss, for each crack, but we do not know the true parameter values, $\boldsymbol{\beta}$. For simulations the responses are random, based on the known values for $\boldsymbol{\beta}$. Since the outcome is random a very large number of realizations is necessary so that we can observe the long-run outcome.

Rather than using likelihoods based on individual hit or miss results, we redefine the loglikelihood to be a weighted average loglikelihood with the weights supplied by the underlying true relationship.

$$\log(L_i) = \log\left(f(a_i, \hat{\boldsymbol{\beta}})\right) \times P(Y = 1 \mid a_i, \boldsymbol{\beta}) \qquad \text{(Eq. 21)}$$

It is important to highlight the distinction between $\boldsymbol{\beta}$, the vector of *known* true model parameters, and $\hat{\boldsymbol{\beta}}$, the *estimates* of the parameter values. The carat symbol (*i.e.* "^") over a parameter is used in statistical practice to indicate that the quantity is an estimate of that parameter and not a known value. In a simulation, $\boldsymbol{\beta}$ is known. In a real experiment $\boldsymbol{\beta}$ is *not* known, so we must rely on $\hat{\boldsymbol{\beta}}$ as a surrogate to glean information about $\boldsymbol{\beta}$.)

4.8.2 Advantages of surrogate Monte Carlo

The dotted confidence lines in Figures 8–14 are based on this method, and agree with the large sample MC simulations. As demonstrated earlier, the surrogate Monte Carlo method produces long-run, asymptotic responses that would otherwise require hundreds of thousands of simulations. Thus we can investigate things like the influence of sample size (number of cracks), size distribution (uniform, "normal", skewed), specimen coverage (the fraction of the POD curve for which the specimens provide useful information; the POD *vs.* size plots in this report indicate coverage with a bold, solid red line) and other studies. Furthermore, these studies are precise. That is, small effects are more easily discernible since they are not obscured by MC randomness. In the following, we summarise the results of our studies.

5 Results

5.1 Effect of sample size on confidence bound width

The surrogate Monte Carlo approach was first applied to investigate the effect of sample size on confidence bound width. The results of this study are shown in Figures 16-23.

Not surprisingly, these figures demonstrate that a larger number of targets produces greater precision in determining a_{50} and a_{90}, as measured by their respective confidence bounds, $a_{50/95}$, and $a_{90/95}$. Doubling the number of specimen, from N=30 to N=60, effectively halves the width of the confidence interval at POD = 90%. Doubling the number of targets again to N=120, however, produces a much less dramatic improvement. The increase in precision diminishes as number of samples increases, so that after about N=90 the addition of more samples does relatively little to further the improvement in precision.

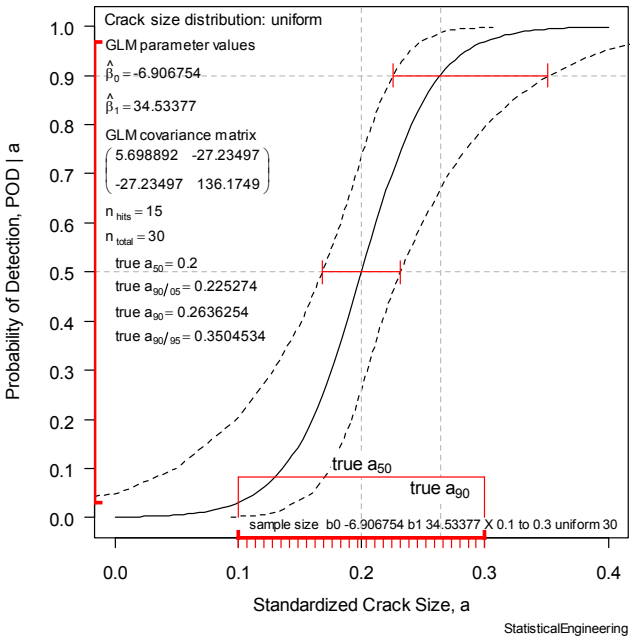


Figure 16
30 targets, uniformly distributed between 0.1 and 0.3, centred on the true (but unknown) POD vs. size relationship.

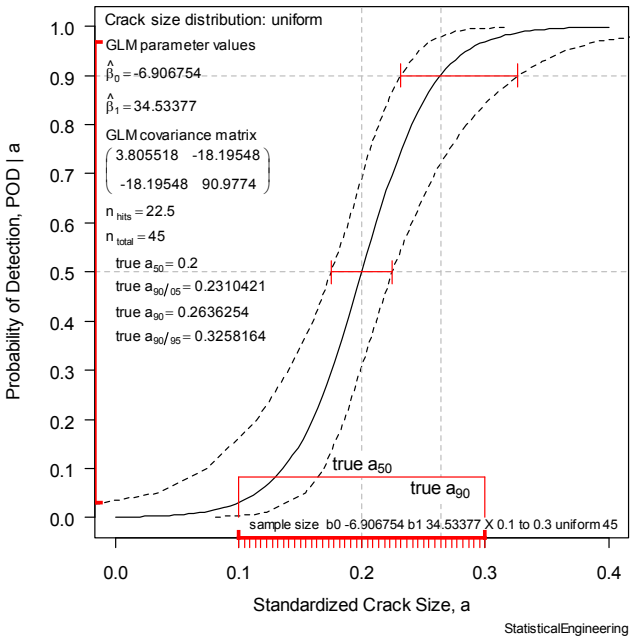


Figure 17
45 targets, uniformly distributed on 0.1, 0.3.

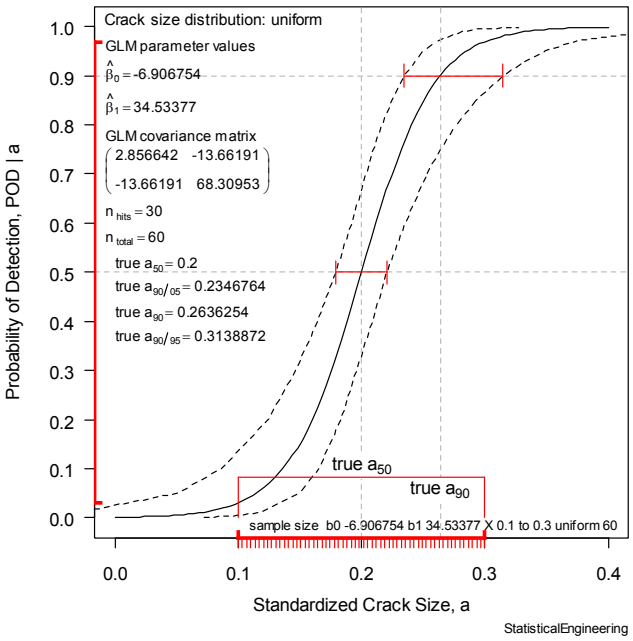


Figure 18
Baseline configuration: 60 targets, uniformly distributed between 0.1 and 0.3, centred on the true (but unknown) POD vs. size relationship

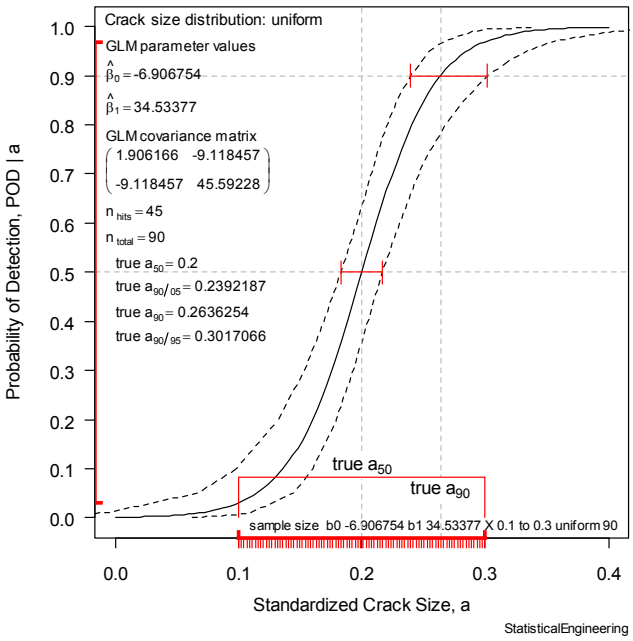


Figure 19
90 targets, uniformly distributed on 0.1, 0.3.

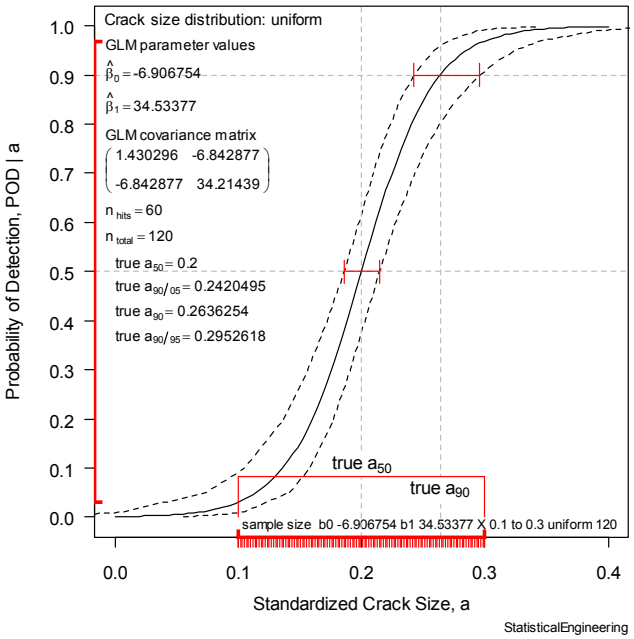


Figure 20
120 targets, uniformly distributed on 0.1, 0.3..

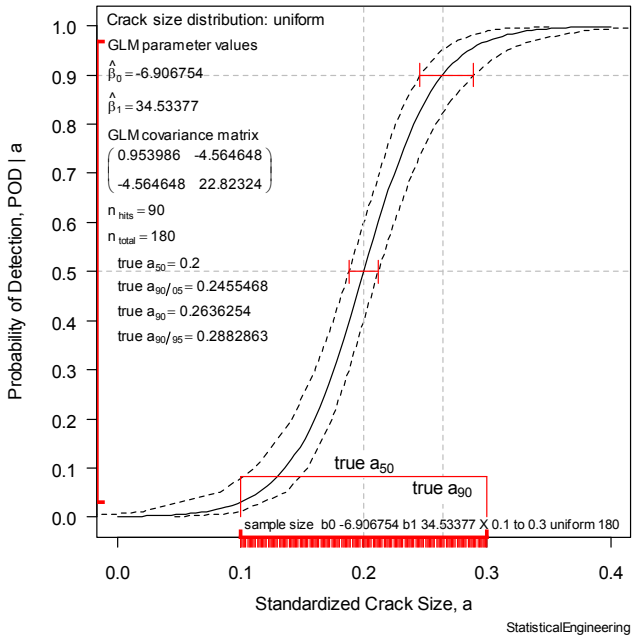


Figure 21
180 targets, uniformly distributed on 0.1, 0.3.

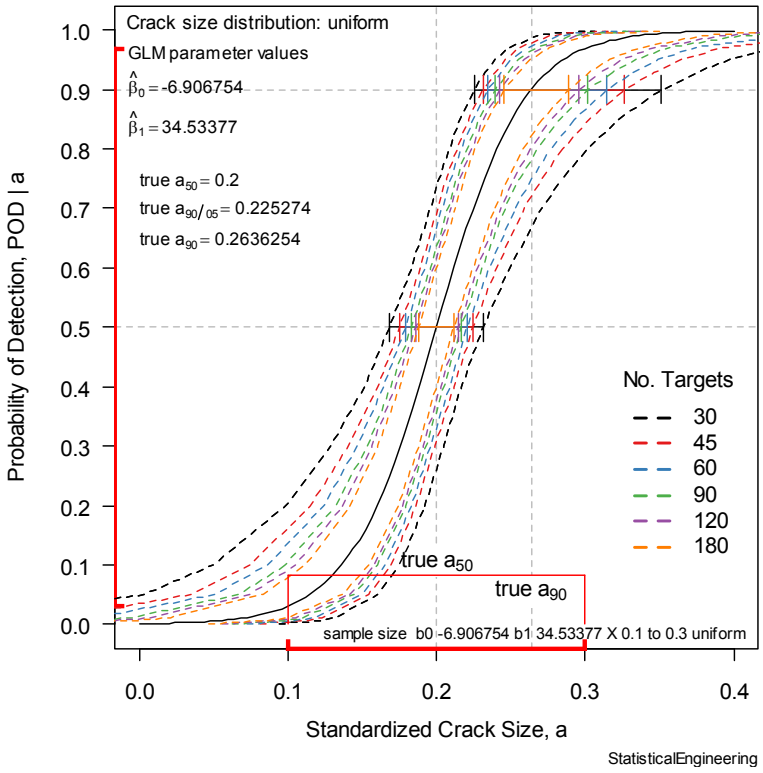


Figure 22
POD vs. size curves showing confidence width decreases with increasing sample size

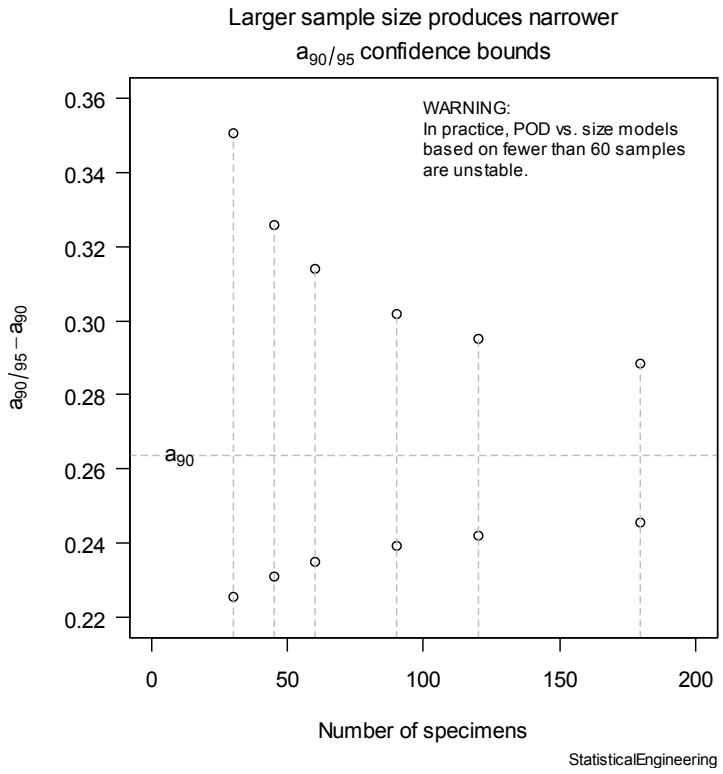


Figure 23
Summary of Figure 22, showing width of 95% confidence bounds vs. sample size

Figures 22 and 23 summarize the influence of sample size on the width of the 95% confidence bounds at POD=90%.

It should be remembered that logistic models of POD *vs.* size that are based on fewer than 60 observations are not stable. The curves here do not rely on maximum likelihood fitting to determine estimates of the model parameters because our method is based on the long run expected response (a weighted probability), rather than a large number of Monte Carlo simulations of binary outcomes. Thus we can compute expected behaviour for very small number of samples, while in practice many such samples would be incapable of producing credible maximum likelihood parameter estimates. This statement is substantiated by more than three decades engineering experience with real POD data, and by our simulation studies for this report.

5.2 Effect of target size coverage on confidence bound width

We proceeded to investigate what is the optimum range of target sizes for a POD study.

We started by considering the likelihood function for a logistic regression model, illustrated graphically in Figure 24. The y-axis on the left side of the plot represents POD. The POD coverage for $-3 \le X \le 3$ is about POD=0.047 to POD =0.953.

The y-axis on the right side of the plot shows the contributions to the log(likelihood) function. For our purposes the maximum value of the likelihood is one, so that the maximum value of the log(likelihood) is zero. The contribution of a hit from a large target (large *X*) is almost zero. In other words, a specimen with a large crack that was found tells us very little about the inspection capability. We expected to find it, and we did. The contribution of a hit affects

the likelihood dramatically (in the negative direction, since all our likelihoods are less than one) for smaller and smaller targets. This happens because we expect to miss smaller targets, so a hit is more unexpected and its contribution is therefore large. The fitting algorithm will try to move the POD *vs.* size curve to make that hit less unexpected. The influence of an unexpected miss is also plotted in Figure 24.

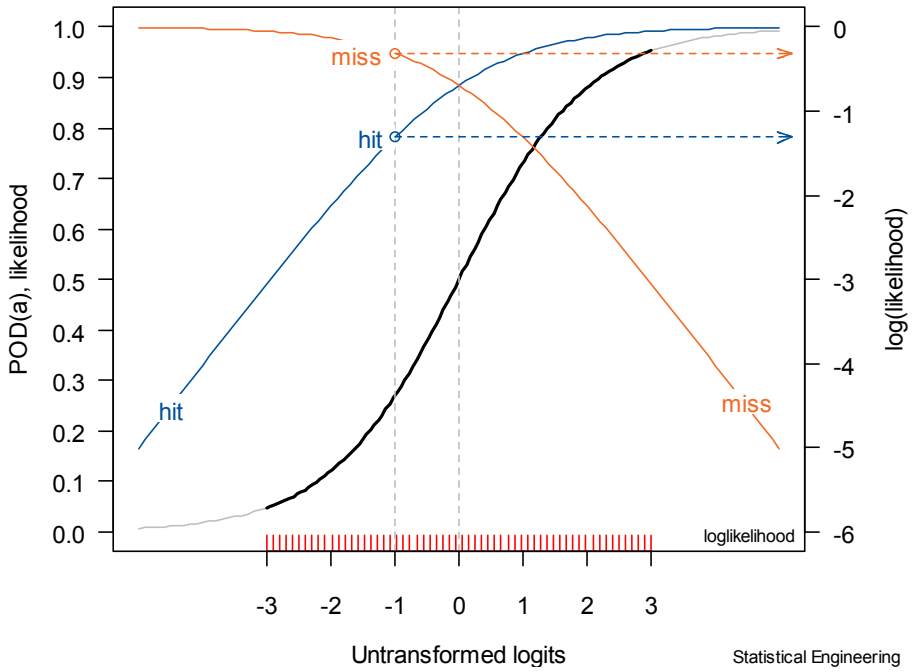


Figure 24
Generic POD vs. size plot showing contributions to the loglikelihood function by a hit and miss.

The following POD *vs.* size plots show how the confidence bound width is influenced by size coverage. These were obtained for a fixed (N=60) sample size, but locating the N targets in progressively wider intervals (widths of 0.05, 0.1, 0.15, 0.2, 0.3, and 0.35) centred near the true POD (but, in practice, unknown) centre (a_{50}=0.2).

In Figure 25, for instance, the N=60 targets are all placed in a very narrow range. Not surprisingly this results in very wide confidence bounds at both low and high POD (a_{90} and a_{10}). What might be surprising, however, is that the bounds at a_{50} are wider than for a slightly wider range of sizes (Figures 26 and 27) even though most of the targets are centred near a_{50}. This is because the parameter values are influenced by all the data, albeit in unequal amounts, as illustrated in Figure 24, so that data missing in the extremes are unable to influence the confidence bounds there.

Our primary interest is in the width of the confidence bounds at a_{90}, so we will concentrate there. As the figures show, increasing the width of size coverage also increases the portion of the POD axis being covered (solid red line on the y-axis of Figures 25–30). As more of the POD extremes (POD > 0.9 or POD < 0.1) are influenced by the size distribution, the narrower the confidence bounds in those regions become.

Nonetheless, increasing size coverage outside corresponding POD range of about 0.03 to 0.97 begins to degrade a_{90} coverage width, and these influences are summarized in Figure

31. This is the result of decreasing effectiveness of observations in the extremes of size (and therefore extremes of POD).

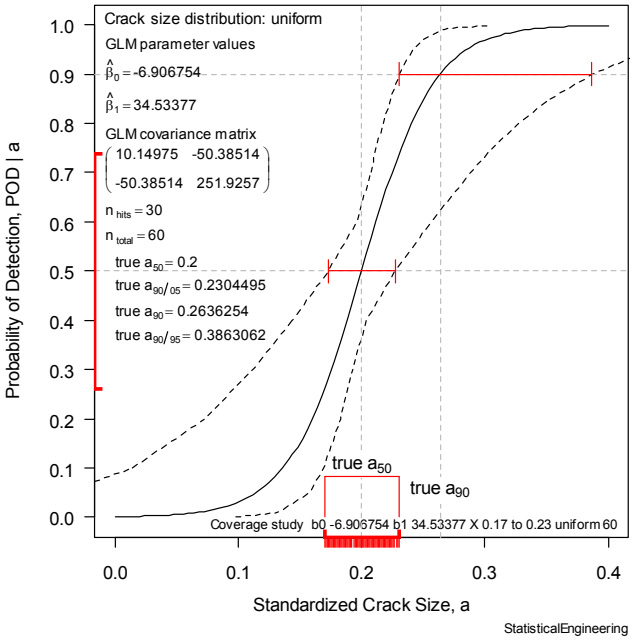


Figure 25
Narrow size distribution produces wide bounds.

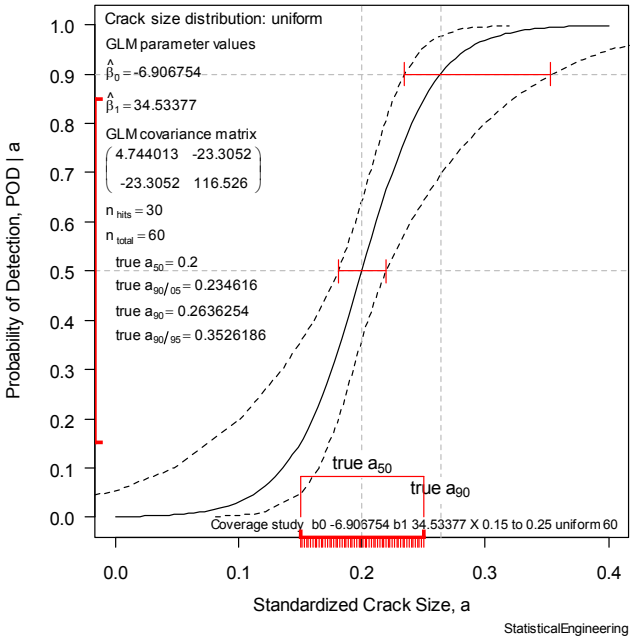


Figure 26
Narrow size distribution produces wide bounds.

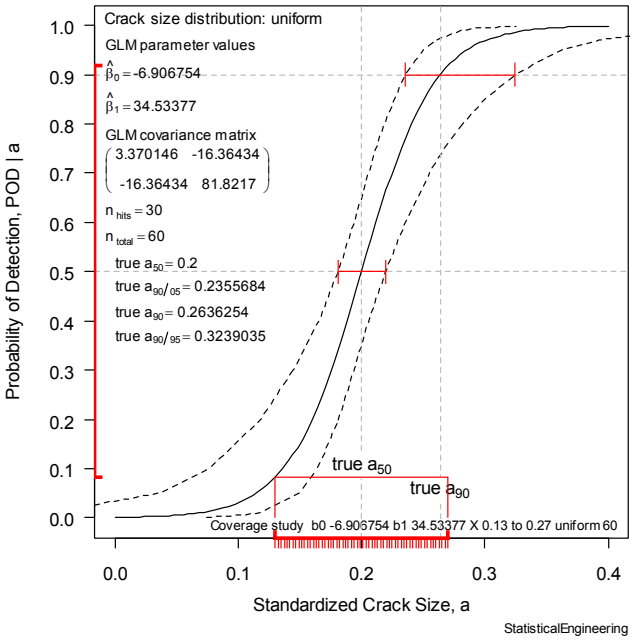


Figure 27
Narrow size distribution produces wide bounds.

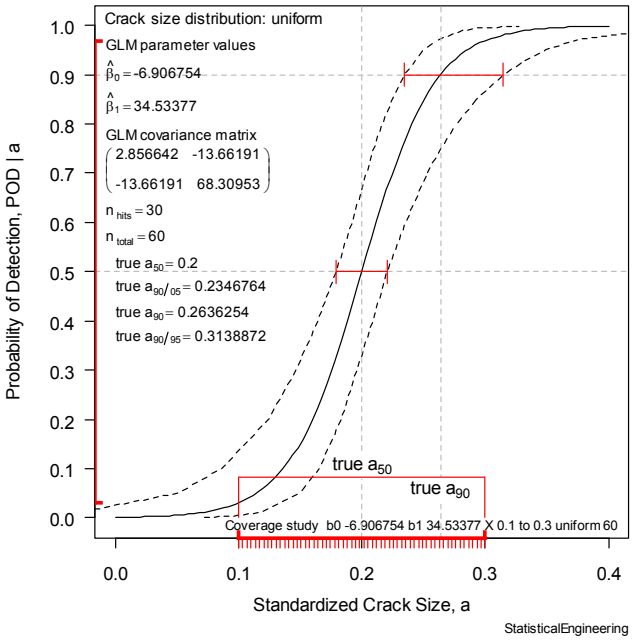


Figure 28
Optimum size distribution covers POD = 0.3 to 0.97.

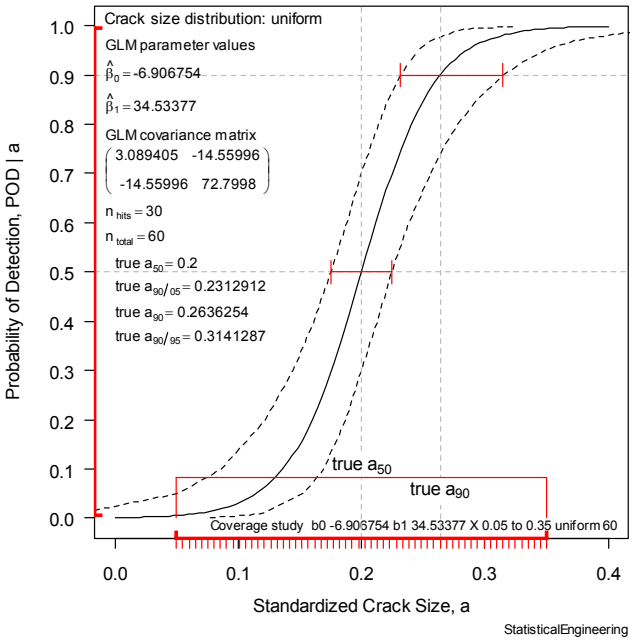


Figure 29
Increasing width past optimum increase bound width.

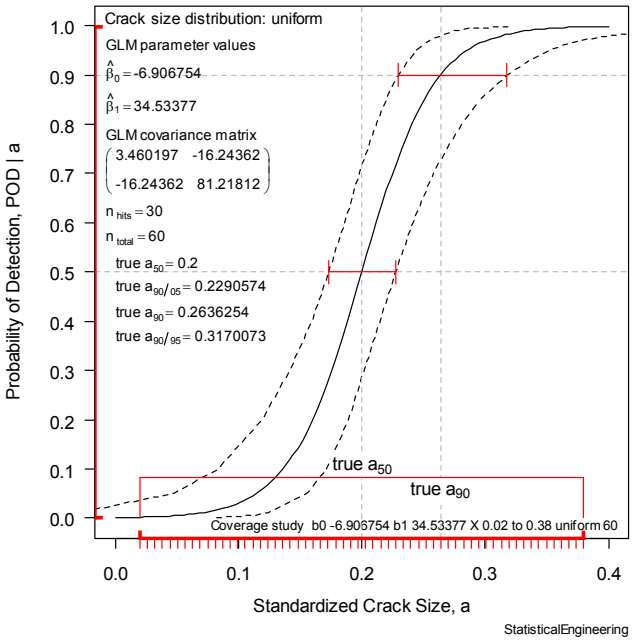


Figure 30
Increasing width past optimum increase bound width.

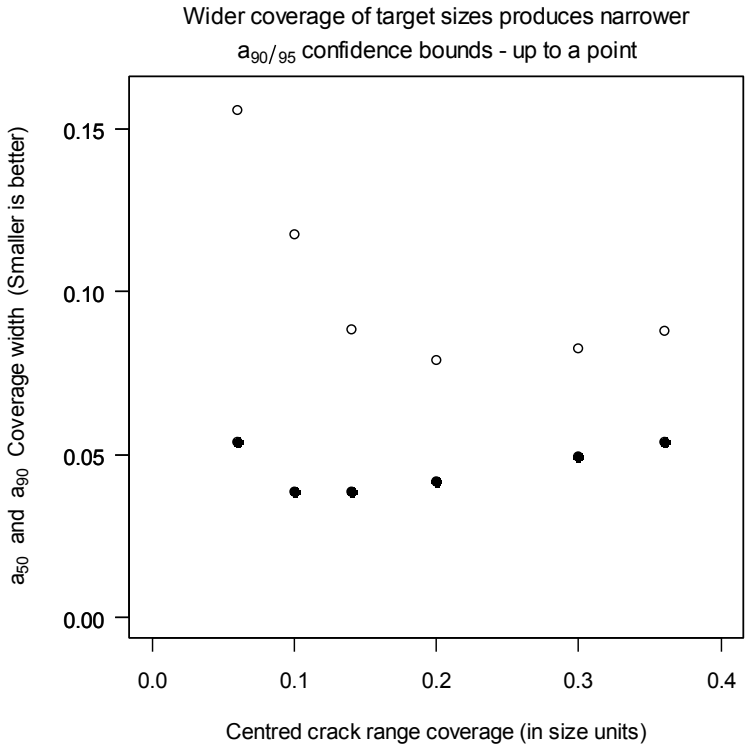


Figure 31
Increasing size coverage outside corresponding POD range of 0.03 to 0.97 begins to degrade a_{90} coverage width.

Figure 31 suggests that the size range from 0.1 to 0.3 in our generic study is optimum. This corresponds to a POD coverage of about POD=0.031 to POD=0.969. If the size range is wider than that, the contributions of those specimens at either end of the range are diminished, leading to an effective decrease in the number of samples. Figures 25–30 show both the size coverage and the POD coverage as solid red lines on the x-axis and y-axis, respectively. In practice, the POD coverage is not known *a priori*. In the face of this uncertainty, Figure 31 suggests that the consequences of overestimating the required coverage are less severe than the consequences of underestimating it.

5.3 Effects of mis-located targets

The foregoing investigated situations where the range of sizes was centred on the centre of the POD *vs.* size curve (which, in practice, is not known *a priori*). It is interesting to consider the effects of mis-location to understand the relationship between coverage on the size axis and the corresponding coverage on the POD axis, as illustrated in Figure 32.

The information in an individual specimen is a function of its location with respect to the true POD *vs.* size curve, as shown in Figure 24. That means that the target sizes must cover most of the POD range to be effective in parameter estimation. Figure 32 provides a mapping of target size coverage to POD coverage.

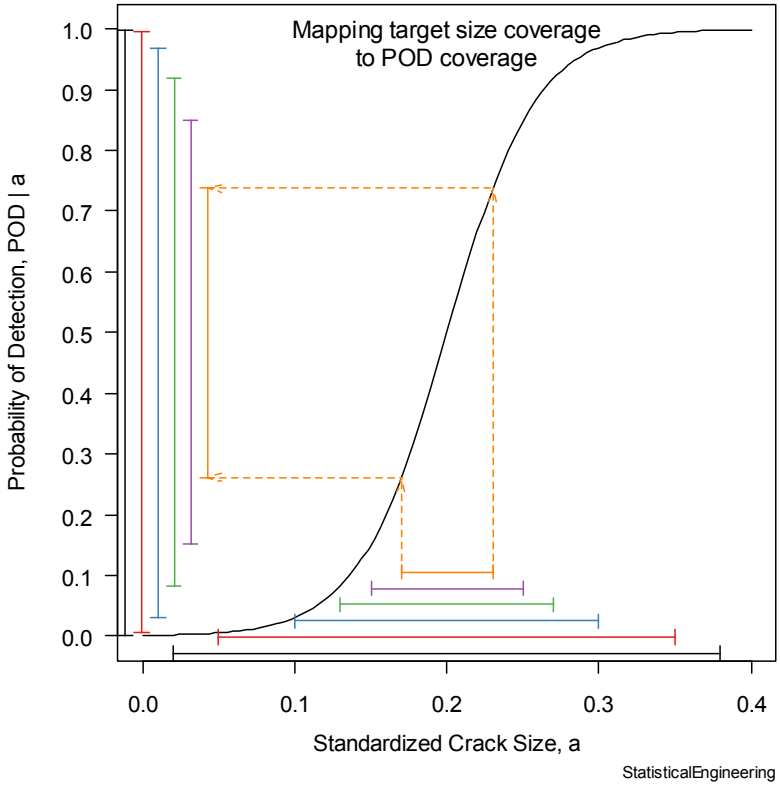


Figure 32
The effectiveness of specimen size depends on how much of the POD range is influenced.

The following figures (33–41) illustrate that targets that are not centred on the true a_{50} adversely influence the width and location of the POD *vs.* size confidence bounds. The effects are not hard to anticipate based on our earlier look at the influence of target sample size, *N,* because mis-locating the target distribution has the effect of decreasing the number

of influential samples (see also Figure 24, Generic POD vs. size plot showing contributions to the loglikelihood function by a hit and miss). These plots were again obtained for N=60.

Figure 33 demonstrates that having all the targets to the left of the true a_{50} produces wide confidence bounds to the right, at a_{90}. This type of mis-location does not, however, have a significant influence at a_{10} because the effective number of specimens has decreased. This is easy to see by noticing that only half of the POD scale is being directly influenced by the targets. The POD coverage provided by the size distribution coverage is shown in Figures 33–41 as the bold red line on the y-axis.

Figures 34–36 show that the effect is the same for targets mislocated to the left by increasingly smaller shifts, but the magnitude of the influences diminishes as the shift diminishes. Figure 37 shows the targets centred at the true a_{50}.

Because of the rotational symmetry of the POD curve, the effects of moving the target size distribution off-centre to the right are analogous to those of mis-positioning to the left, as seen in Figures 38–41. Figure 41 also illustrates that diminishing the effective number of samples by mis-positioning has minimal influence on the confidence bounds at a_{90}, however the bounds at a_{50} are substantially larger than those in Figure 37.

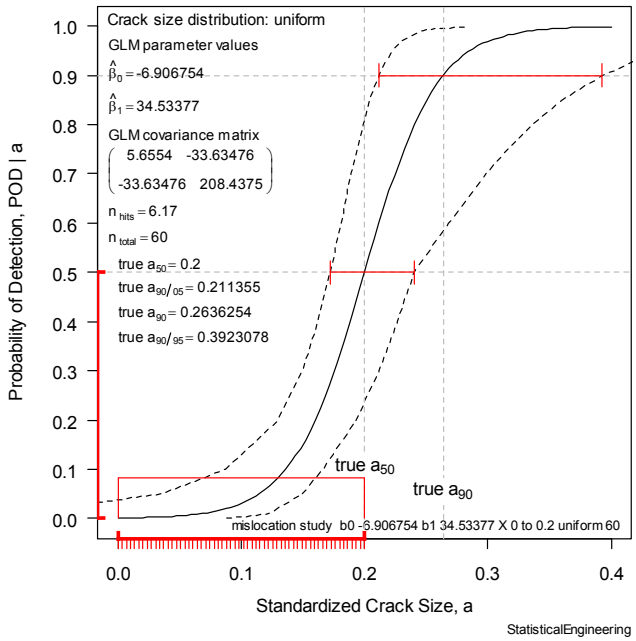


Figure 33
Size distribution misplaced left, centred at 0.1

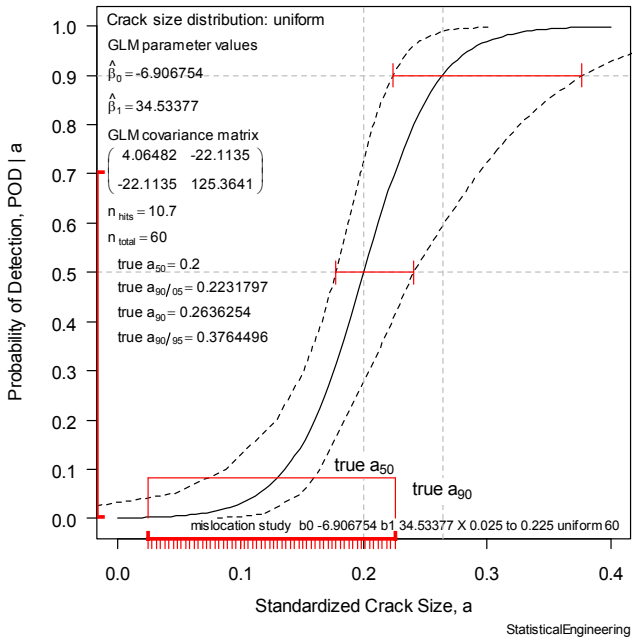


Figure 34
Size distribution misplaced left, centred at 0.125

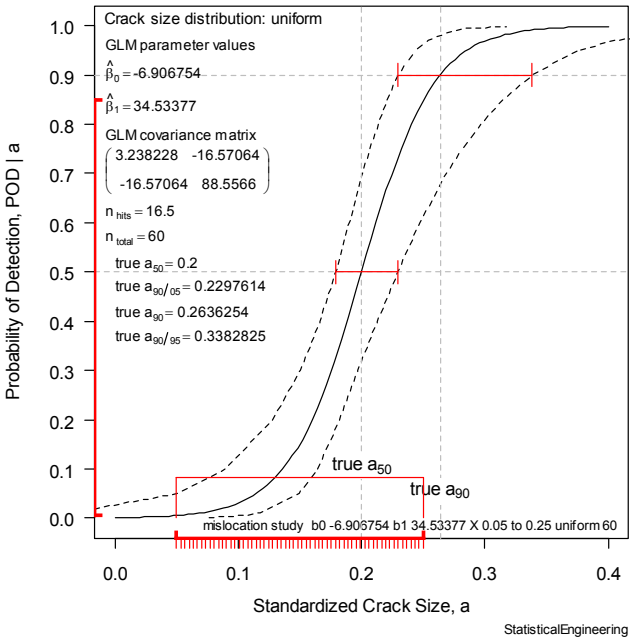


Figure 35
Size distribution misplaced left, centred at 0.15

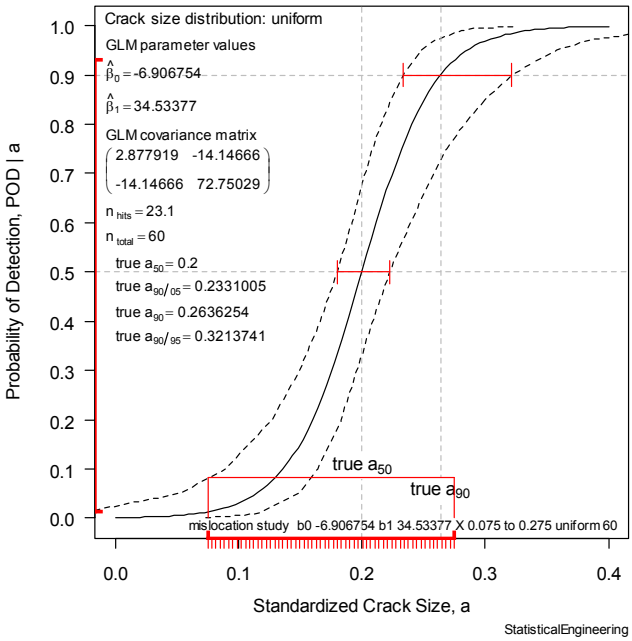


Figure 36
Size distribution misplaced left, centred at 0.175

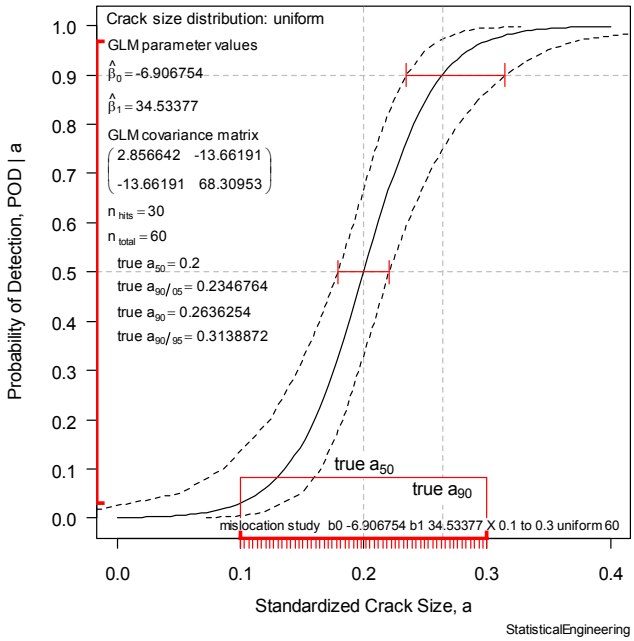


Figure 37
Size distribution, centred on true a_{50}

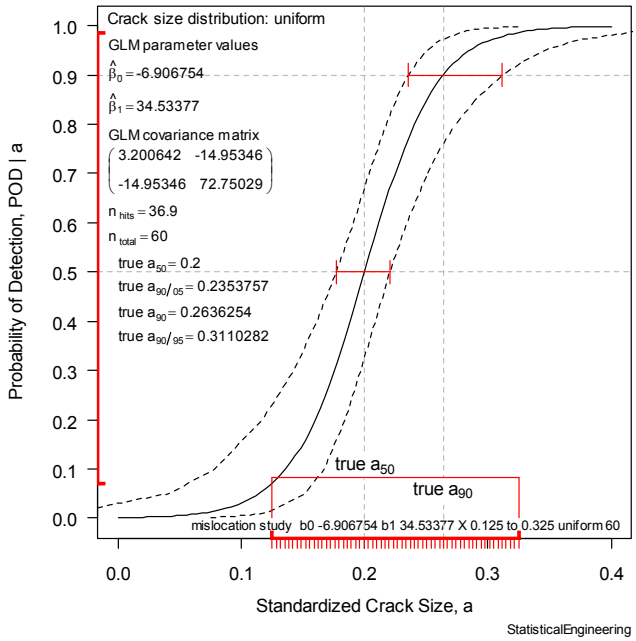


Figure 38
Size distribution misplaced right, centred at 0.225

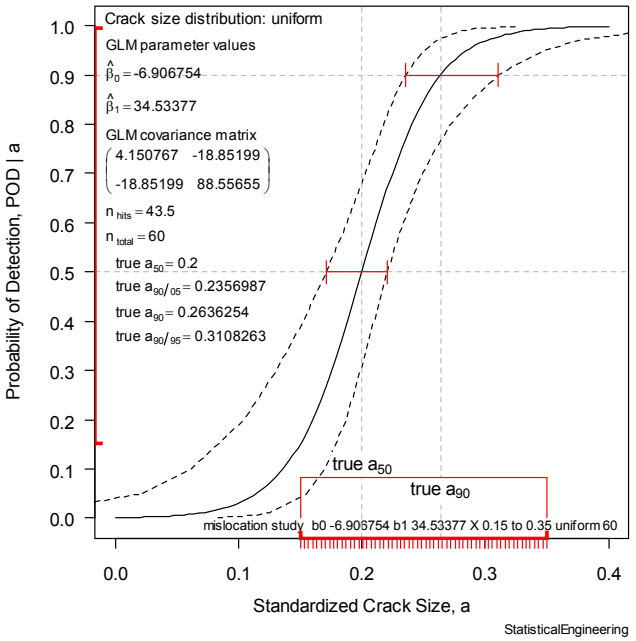


Figure 39
Size distribution misplaced right, centred at 0.25

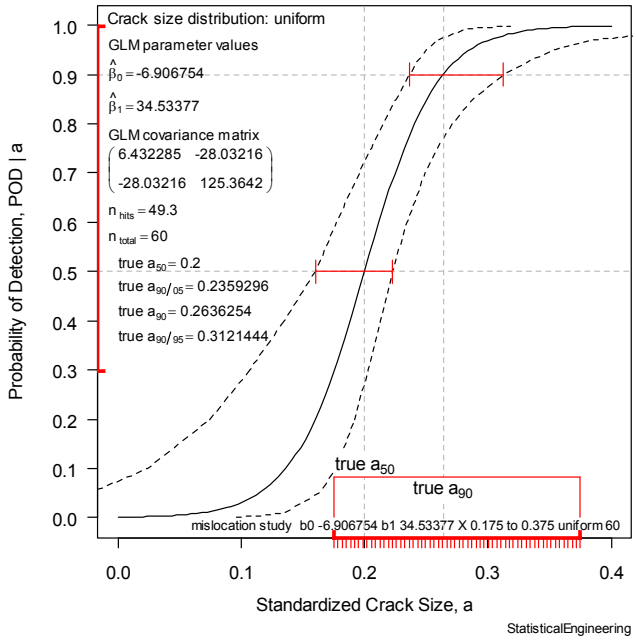


Figure 40
Size distribution misplaced right, centred at 0.275

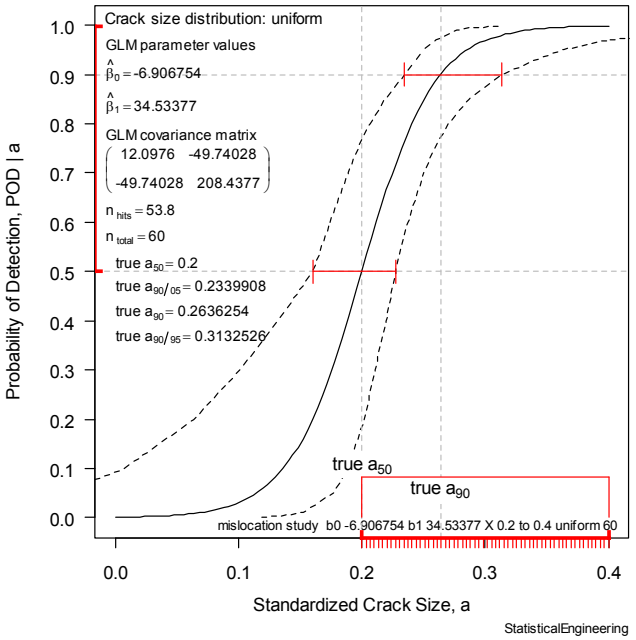


Figure 41
Size distribution misplaced right, centred at 0.3

5.4 Non-uniform size distributions

Figures 42–45 show the influence of the shape of the target size distribution on the POD confidence bounds. Not surprisingly, grouping the targets near the true (but, in practice, unknown) value of a_{90} curve results in the narrowest bounds there, as seen in Figure 43. Grouping the targets near a_{90} with a left-skewed distribution, Figure 44, produces the

narrowest bounds at a_{90}, and thus the smallest $a_{90/95}$. Figure 45, the right-skewed distribution places the fewest targets near a_{90}, and thus results in the widest bounds and the largest ("worst") $a_{90/95}$. These results are summarized as a bar chart in Figure 46.

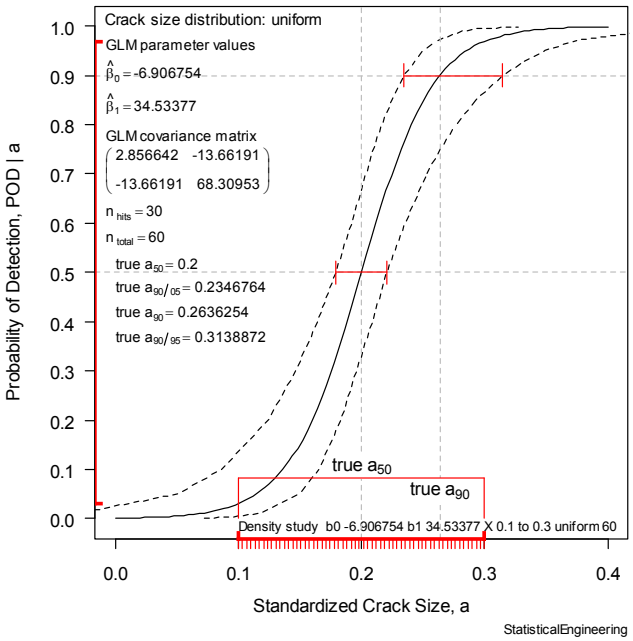


Figure 42
Uniform distribution of target sizes.

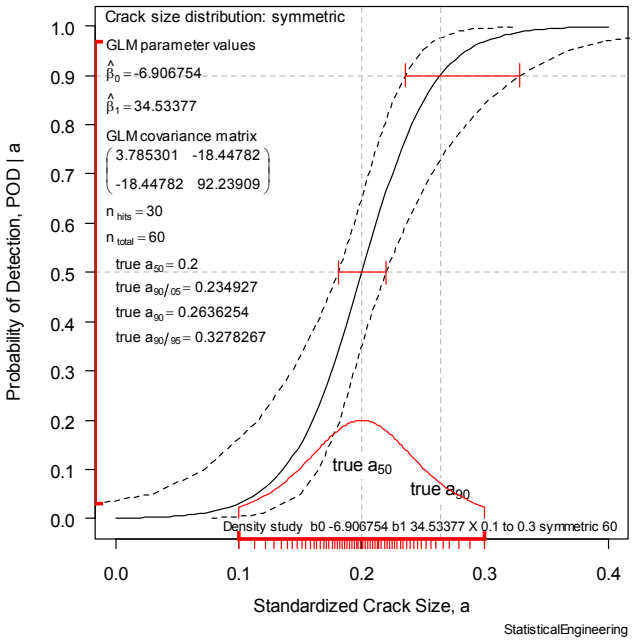


Figure 43
Symmetrical "normal" distribution of target sizes.

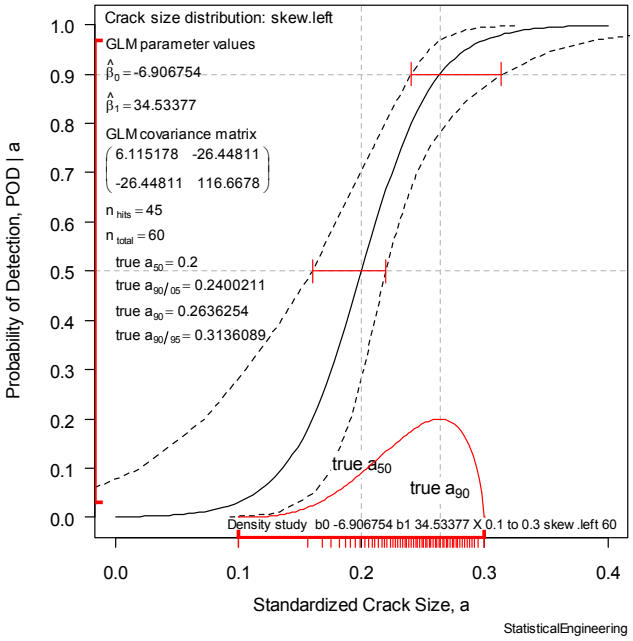


Figure 44
Left-skewed distribution of target sizes.

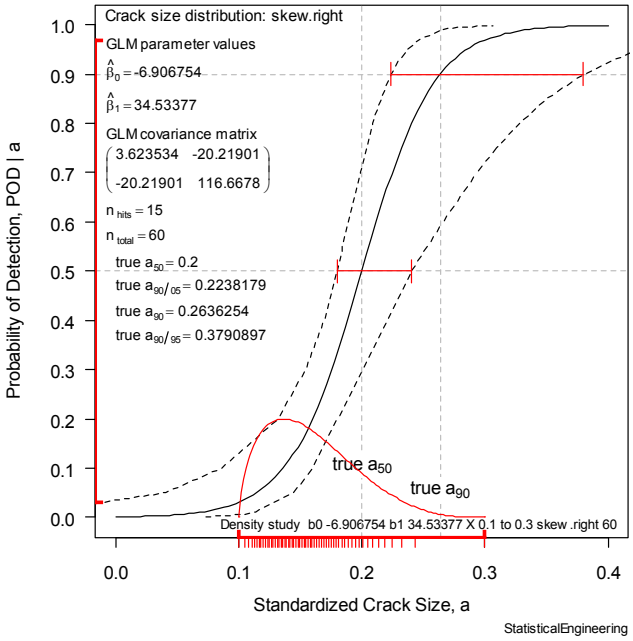


Figure 45
Right-skewed distribution of target sizes.

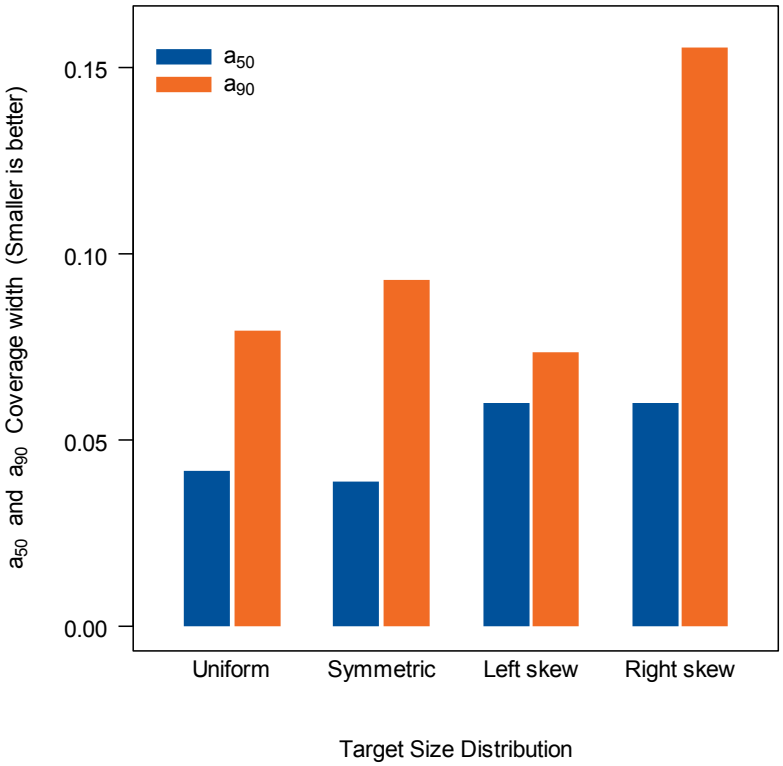


Figure 46
Right-skew has widest bounds at a_{90}.

Considering such outcome, the question arises as to why anyone would consider anything other than a uniform or symmetrical distribution of target sizes.

All of our discussions have been based on placing the specimens with respect to a symmetrical POD *vs.* size curve whose location, while unknown, can be reasonably deduced (at least approximately) by a knowledgeable NDE practitioner, based on his experience with similar inspections. On the other hand, it is true that many POD *vs.* size relationships are only symmetrical with respect to a transformed size, for example log(size)[3].

Figure 47 shows a symmetric POD *vs.* size curve with respect to a log(size) axis. As stated earlier, transformation, scaling or offset has no direct influence on the Generalized Linear Model of POD as a function of size. However, transformations such as log(size) can have a significant *indirect* influence by changing the relationship of the distribution of target sizes with respect to the (unknown) true POD *vs.* size.

[3] Not all POD *vs.* size curves are symmetric or can be transformed into symmetry. Examples are the loglog link and complementary loglog link used by generalized linear models. While these links are used infrequently, they should be considered when transformations on size are not satisfactory. See MIL-HDBK-1823A for further details.

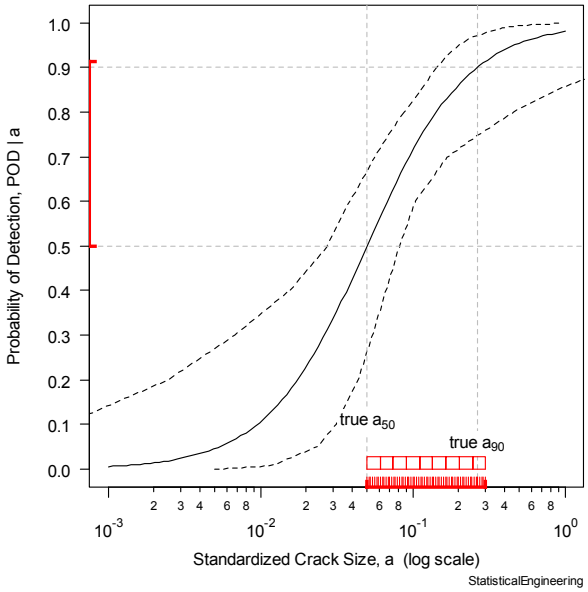


Figure 47
POD vs. untransformed size.

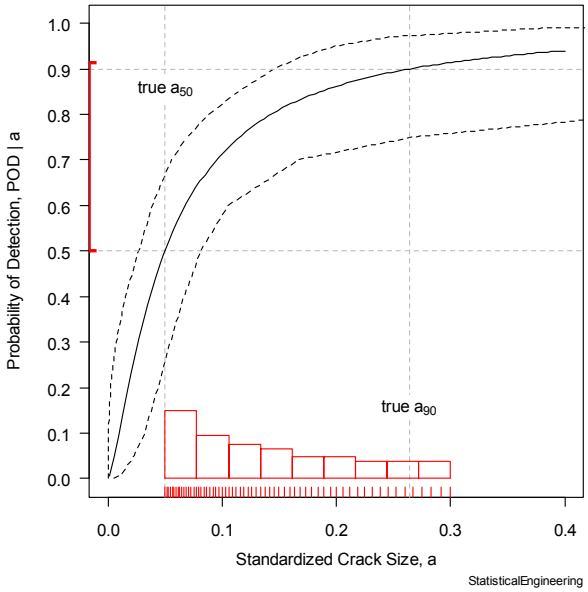


Figure 48
POD vs. transformed size.

Figures 47 and 48 have the same a_{90} as our earlier baseline curves, a_{90} = -1.33322 in Figure 47 and a_{90} = exp(-1.33322) = 0.2636 in Figure 48. Figure 48 is Figure 47 re-plotted on a Cartesian x-axis. The resulting plot no longer has the familiar "*S*" shape. The mathematical description of Figures 47 and 48 is given in (Eq. 6), with *h(a)*=log(a) for both Figures 47 and 48. Thus these curves are mathematically identical: their only difference is in how they are plotted.

Notice that the uniform distribution of log target sizes in Figure 47 has now become a right-skewed distribution of transformed sizes in Figure 48. Notice also that even though the range of coverage in the figures is from size = 0.05 to 0.3, which is wider than our baseline range of 0.1 to 0.3, the resulting POD coverage is only POD = 0.5 to 0.91, and our earlier discussion concluded the most effective coverage should be from POD = 0.03 to 0.97.

The difficulty in practice is that the true shape and location of the POD *vs.* size curve is unknown *a priori,* but if it is suspected (e.g. from previous experience or theoretical insight) that the underlying relationship of POD is not directly linear with size but with log(size), or some other similar transformation, then a right-skewed distribution of sizes would afford some protection against ineffective specimen size allocation.

As the histogram of sizes in Figure 48 suggests, there is not a simple probability distribution that can be used to model the transformed sizes that is useful in all cases. That is because the degree of nonlinearity between *a* and *h(a)* depends on two factors: (1) the range of sizes, and (2) the proximity of the smallest size to zero. To help the practitioner in designing a POD *vs.* size study the Guidelines (section 6.2) presents a worksheet for allocating experimental resources to determine how many of what size targets should be fabricated.

Figure 48 also illustrates the dangers of extrapolating a mathematical model beyond the range of the data supporting it. In the figure, the range of support ends at POD = 0.5. That means that the convergence of confidence bounds at *a* = 0 is an artefact of the log(size)

model (since log(0) is undefined) and less likely due to observable behaviour. This is more evident in Figure 48, showing divergent confidence bounds as size approaches zero. The limits of extrapolation are too often overlooked, especially if there is no indication on the plot of what is extrapolated.

Figure 49 compares the POD curves of Figure 4 and Figure 48 and reveals that they share the same a_{90}, but a_{50} = 0.2 for Figure 4 is four times larger than a_{50} =0.05 for Figure 48. Does Figure 48, with the smaller a_{50}, represent the better inspection because it can find smaller targets? Probably not, but it would take an ancillary study of Probability of False Positive (PFP) to confirm that. Parallel studies of POD and PFP are recommended best practices in MIL-HDBK-1823A and in Gandossi and Annis (2010).

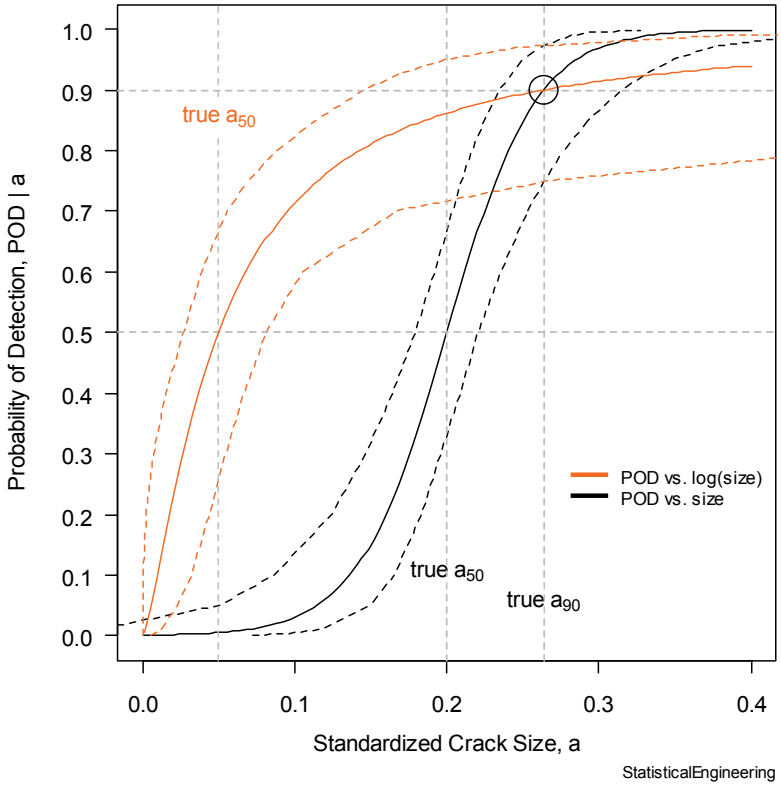


Figure 49
Inspection with the smaller a_{50} is inferior.

The inspection capability for targets larger than a_{90} demonstrates that the inspection described by the curve in Figure 4 is superior in two regards: (1) it has a higher POD for targets larger than a_{90}, and (2) it has a much narrower confidence bound at that crack size. What differentiates a good inspection from a less effective one usually depends on how rapidly the POD changes from smaller (innocuous) sizes to larger (pernicious) ones. Thus finding smaller sizes is not helpful if the inspection is unable to discriminate tolerable defects from those that are not. Finally, Figure 49 illustrates why comparing POD *vs.* size curves on the basis of a single point, like a_{90}, is very misleading.

How can it be decided if the unknown POD *vs.* size relationship is more effectively described with a log(size) transform? With real binary data this is straightforward: perform a likelihood ratio test to see which model is better at describing the data. In some instances, theoretical insight or evidence from previous experience can also inform the selection of a suitable

model. Provided that the range of POD coverage is adequate, the **mh1823POD**[4] software can be used to perform this comparison.

Suppose that it is believed that the true POD *vs.* size relationship will require a log(size) transform. How would one determine appropriate sizes for NDE lab specimens? A simple uniform distribution of sizes in log-space has a simple one-to-one transform back to Cartesian space (*viz*. exp(log(X)), however the result depends on the proximity to zero of the smallest size. This is illustrated for *N*=60 in the following figures.

Figure 50 shows three uniform distributions of log target size. Each distribution is 0.2 size units wide. The histogram shows ten intervals, each with the same width. The number of targets in each interval is indicated by the height of the histogram bars and is determined by counting the number of sizes in each interval (6 in this example).

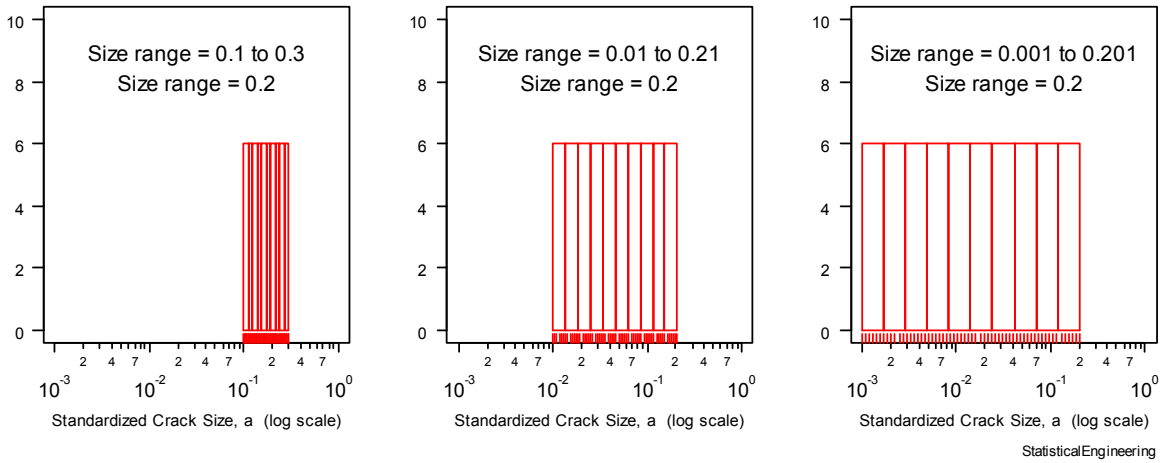


Figure 50
Three size ranges, all 0.2 wide, uniformly distributed in log(X)

Figure 51 plots the sizes on a Cartesian X-axis, again with ten intervals of constant width. The height of the bar for each interval is determined as in Figure 50, by counting the number of sizes in that interval. The histograms are all right-skewed and, although they all are 0.2 units wide with ten intervals, and each interval is 0.02 wide, the numbers in each interval vary considerably. Which one is "best"?

There is no single answer, and in practice the distribution is usually determined empirically, based on experience. The examples is this paper augment that experience by illustrating the effects of having too wide or too narrow a distribution of sizes, the influence of not centring the distribution on the true (but unknown) POD *vs.* size relationship, and the effects of distributions other than uniform (*e.g.* symmetric (“normal”), left- and right-skewed. A sketch of the desired distribution of sizes can then be converted into a histogram and thus the number of targets in each size range can be determined. We present a refinement of the “sketch” method in the Guidelines.

[4] One of the authors of this report (C. Annis) has written a software add-on (**mh1823 POD**) to carry out POD analyses, applying the methods presented in MIL-HDBK-1823A (2009). The add-on has been developed for **R**, the emerging world standard for statistical computing and graphics**. R** is a *free* software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS (see http://www.r-project.org/). The software add-on **mh1823 POD**, that works on Windows and Windows emulators, is available *for free* at http://mh1823.com/mh1823.

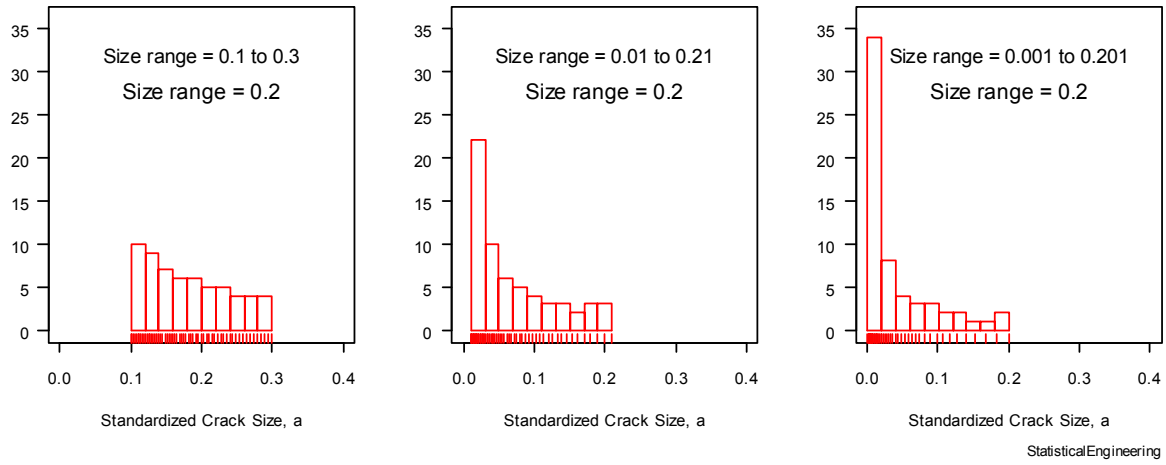


Figure 51
The degree of skew for a log(X) transform depends on proximity of X_{min} to zero.

6 Summary and conclusions: guidelines for practitioners

The primary objective of this work was to provide guidelines for NDE practitioners for designing experiments to assess the effectiveness of a binary response inspection system using POD *vs.* size curves. A second, related, objective was to substantiate these guidelines through Monte Carlo studies. As such this work expands on our earlier effort, Gandossi and Annis (2010), "Probability of Detection Curves: Statistical Best-Practices, ENIQ report No 41.

Before proceeding, we remind the reader of the scope of applicability of MIL-HDBK-1823A methods. It is often taken for granted that NDE practitioners are well aware of the following criteria, but experience has shown that this may not always be the case.

6.1 Reminder of the Scope of Applicability of MIL-HDBK-1823A Hit/Miss Methods

1. The specimens must have targets with measurable characteristics, like size or chemical composition. This precludes amorphous targets like corrosion unless a specific measure can be associated with it, such that other corrosion having that same measure will produce essentially the same output from the NDE equipment.

2. The mh1823 POD software assumes that the input data are correct. That is, if the size is X, then that is the true size. If the response is Y, then that is the true response. Situations where these conditions cannot be ensured (*e.g.* where target sizing is only approximate) will necessarily result in only approximate results. (The problem of accurate crack sizing is discussed in MIL-HDBK-1823A, Appendix I.1 "Departures from Underlying Assumptions – Crack Sizing and POD Analysis of Images.")

3. The MIL-HDBK-1823A statistical methods used in this paper require that a POD curve goes to zero on the left, and to one on the right. These conditions are easily met by most, BUT NOT ALL, POD data. Data for which min(POD)>0 (perhaps due to signal contamination by excessive background noise), or max(POD)<1 (resulting from random misses not related to target size) cannot be correctly represented by a model for which min(POD)=0 and max(POD)=1. This is illustrated in Figure 52. (See MIL-HDBK-1823, Appendix I.4 "Asymptotic POD Functions"). As discussed earlier and

illustrated in Figure 24, specimens with large targets that will almost surely be found, or very small targets that will almost surely be missed don't contribute significantly to the likelihood function. But if there is a POD "floor" or POD "ceiling" then these specimens would be very important because without them the existence of these asymptotes would go undetected.

How can a practitioner tell if the model agrees with the data? One way is to look at the POD *vs.* size plot. If there are misses at large sizes where the POD is near 1, or hits at small sizes where the POD is near 0, then the model does **not** agree with the data. Whenever **any** model disagrees with the data being modelled, the result will be **wrong**. The software may be coaxed into producing a POD(a) curve, but it will be a **wrong** curve, sometimes with $a_{90/95}$ values where none exist. Further guidance on assessing the adequacy of logistic regression models, including formal tests of goodness-of-fit, can be found in Chapter 5 of Hosmer and Lemeshow (1989).

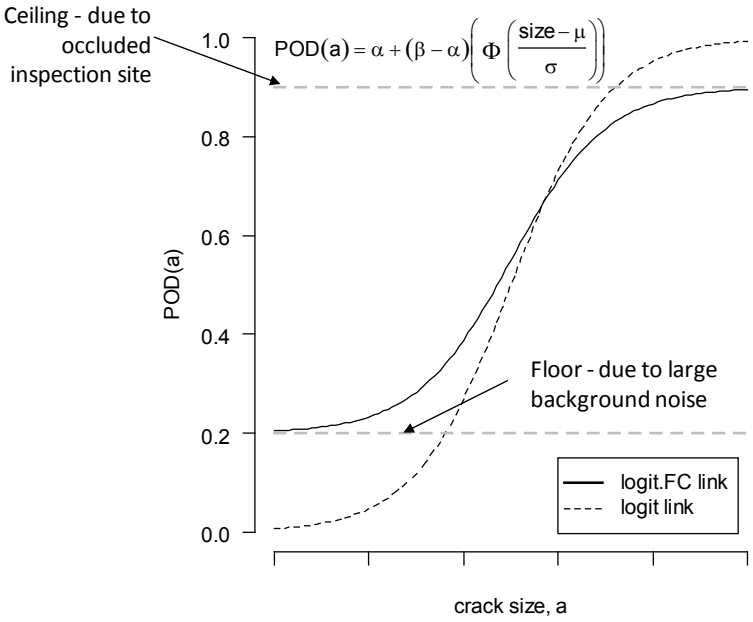


Figure 52
The methods in this paper are valid only for inspections for which POD_{min}= 0 and POD_{max}= 1.

6.2 Guidelines

1. The recommended minimum number of targets for hit/miss POD *vs.* size modelling is *N*=60.

2. A uniform distribution (after transformation where appropriate) of target sizes is recommended.

3. The target range should result in POD coverage from about POD=3% to POD = 97%. (See Figure 32.)

4. Fewer than 60 targets **should not** be used:

    a. Using as few as 30 or 45 specimens occasionally results in numerically unstable situations with non-convergent GLM parameter estimation, or convergence to nonsense parameter values, for example POD that gets worse as size increases.

    b. Even with successful convergence, the confidence bounds are quite wide and so the precision in estimating a_{90} is poor. (See Figure 23.)

5. While N=90 specimens is often worthwhile, increasing the number further produces diminishing returns and is often not cost-effective. (See Figure 23.)

6. If it suspected that the POD *vs.* size model will use a log(size) transform, then a right-skewed distribution of targets is recommended.

    a. As with non-transformed sizes the POD coverage should be about 3% to 97%. If in doubt, it is generally preferable to overestimate the required coverage than to underestimate it. (See Section 5.2.)

    b. Figure 53 illustrates a schematic diagram that may be helpful in determining appropriate size ranges for skewed size distributions.

7. The possibility of the existence of POD “floor” or POD “ceiling” should be kept well in mind:

    a. Not all inspections can achieve 100% POD for very large crack sizes. The existence of a POD ceiling will require special statistical analysis methods.

    b. Some inspections are so confounded with noise, especially for small size targets, that the assumption of $POD_{min} = 0$ is untenable. This situation, too, will require special statistical procedures.

6.3 How to allocate target sizes in lab specimens

Allocating NDE specimens can be more problematic than it might seem, from a practical point of view. This is especially true when the desired crack size distribution is not uniform. Indeed, experience has shown that, when specifying to the machine shop what size cracks are to be placed in the specimen, it is not a good idea to indicate exact crack sizes. It can be quite costly to manufacture an exact crack size and much easier and less expensive to specify a size range.

Thus, we recommend to specify (1) a set of size intervals and (2) how many cracks should be manufactured in each of these intervals. As a starting point, try to place about half of the targets on each side of the anticipated a_{50}. This involves some guesswork, of course, since the location of a_{50} is one of the things to be determined by the POD experiment.

We suggest using an approach as depicted in the schematic diagram of Figure 53. The idea is to draw a series of boxes that can be rearranged to form the desired size distribution.

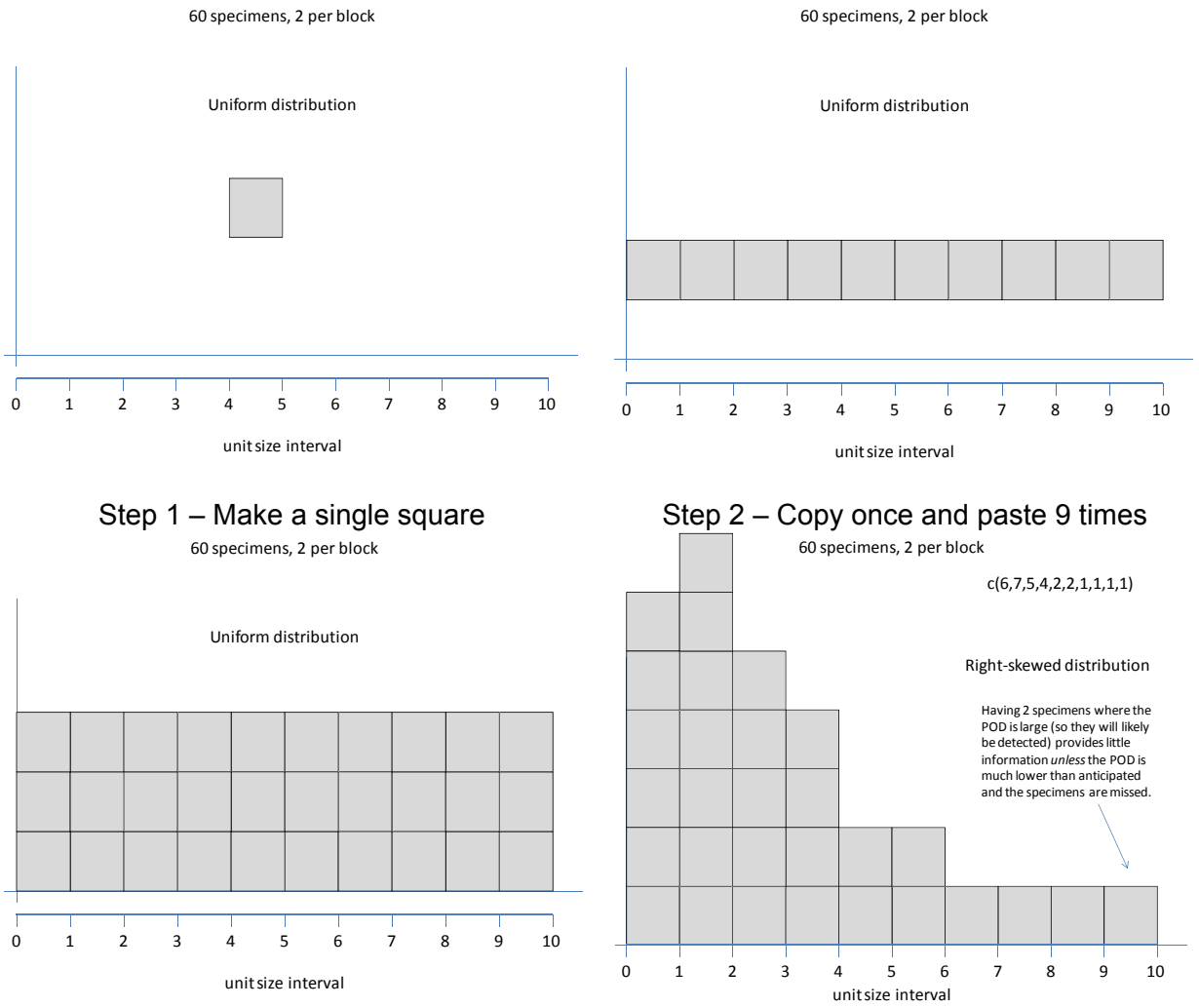
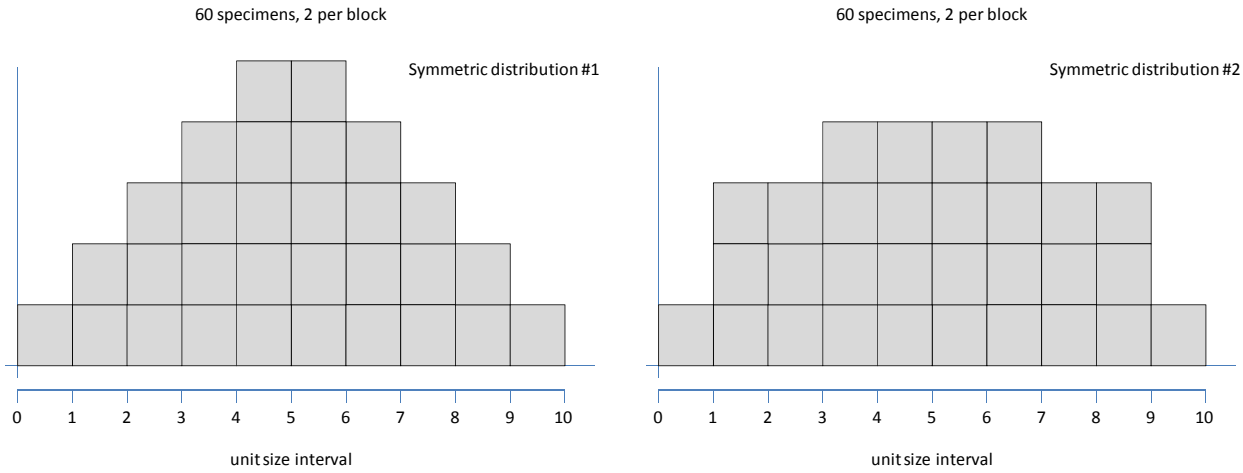


Figure 53
Worksheet for allocating cracks in NDE lab specimens

The following steps describe how to make a worksheet like Figure 53. We assume that N=60.

- Create a size X-axis and a quantity Y-axis, and a single square. Each square will represent 2 lab specimens, thus 30 squares are required.
- Make the square small enough that ten can be placed side-by-side horizontally.
- Next, copy the square and paste it nine times, making ten boxes.
- Place the ten boxes side-by-side together and copy all ten as a single unit.
- Then paste twice, producing 30 boxes in all.
- Experiment by moving the individual boxes to form the desired distribution of sizes (see Figure 53).
- Finally translate the dimensionless boxes into meaningful dimensions by dividing the real size range by 10, and adding the minimum size to each endpoint.

Figure 54 shows two examples of the method applied to symmetrical (“normal”) crack size distributions.



Example: Symmetrical Distribution #1 Example: Symmetrical Distribution #2

Figure 54
Examples of symmetrical ("normal") crack size distributions.

7 Acknowledgements

The authors would like to express their thanks to Charles Schneider (TWI, United Kingdom) and Iikka Virkkunen (Trueflaw, Finland) for reviewing earlier drafts of this paper and for providing very valuable comments.

8 References

Bullough R, Dolby R E, Beardsmore D W, Burdekin F M and Schneider C R A (2007) "The probability of formation and detection of large flaws in welds". TAGSI document TAGSI/P(01)173. International Journal of Pressure Vessels and Piping, Volume 84, Issue 12, Dec, pp730-738. http://dx.doi.org/10.1016/j.ijpvp.2007.06.015.

Gandossi and Annis (2010), ENIQ report No 41: "Probability of Detection Curves: Statistical Best-Practices", EUR – Scientific and Technical Research series – ISSN 1018-5593, ISBN 978-92-79-16105-6. Available for download at http://safelife.jrc.ec.europa.eu/eniq/docs/eur_reports/ENIQ%20report%2041.pdf

Hosmer D W and Lemeshow S (1989), **Applied logistic regression**, John Wiley & Sons, New York.

Marshall W (1982) "An assessment of the integrity of PWR pressure vessels – Section 10". Second Report by a Study Group under the Chairmanship of Dr W Marshall, UKAEA, March.

MIL-HDBK-1823A (2009), "Nondestructive Evaluation System Reliability Assessment," Standardization Order Desk, Building 4D, 700 Roberts Avenue, Philadelphia, PA 19111-5094. Available for download at http://mh1823.com/mh1823.

Meeker and Escobar (1998), **Statistical Methods for Reliability Data**, Wiley, especially Appendix B, "Some Results from Statistical Theory," p 617 ff.

Appendix 1

Our early simulations used odd numbers of targets to avoid having half the “data” on each half of the centre and thus omitting the greatest contributor to the likelihood function, which is the centre point for uniform and symmetrical distributions.

We investigated this matter, and the outcome was that our concern was overwrought. Figure 55 shows that the differences between 60 and 61 targets are quite small, with, for example, a difference in $a_{90/95}$ of 0.0005868 (less than 1% of the width of the 95% confidence band at a_{90}).

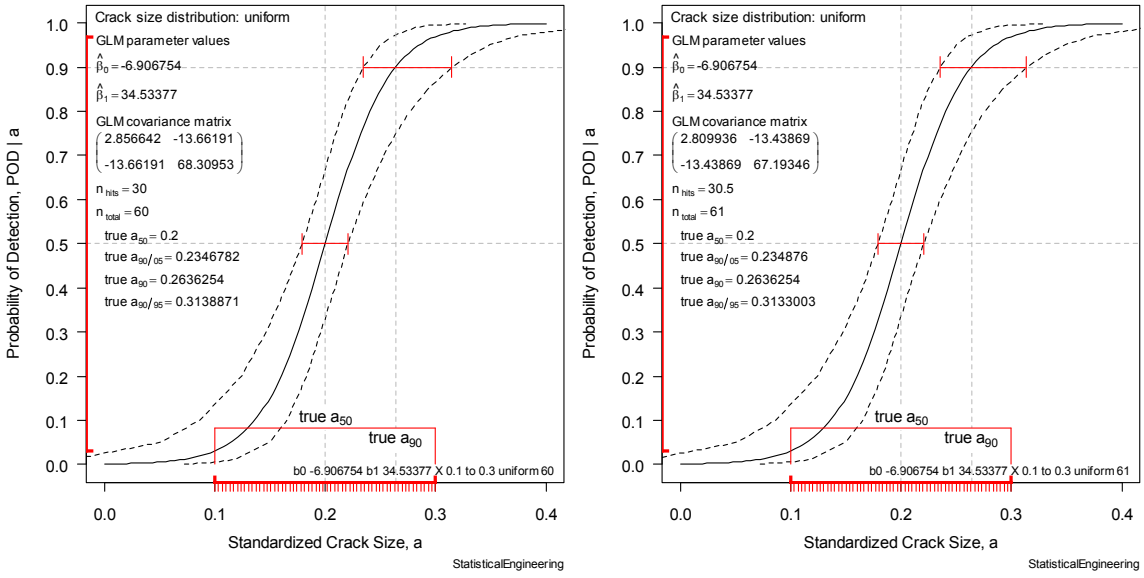


Figure 55
The differences between 60 and 61 specimens are very small

This page is intentionally left blank.

European Commission

EUR 24429 EN – Joint Research Centre – Institute for Energy
Title: ENIQ TGR TECHNICAL DOCUMENT – INFLUENCE OF SAMPLE SIZE AND OTHER FACTORS ON HIT/MISS PROBABILITY OF DETECTION CURVES

Author: Charles ANNIS (Statistical Engineering)
Luca GANDOSSI (DG-JRC-IET)

Luxembourg: Publications Office of the European Union
2012 – 50 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1831-9424
ISBN 978-92-79-23018-9
doi:10.2790/43050

Abstract
The purposes of this document, aimed mostly at NDT engineers and practitioners, are threefold: (1) to extend the conclusions of an earlier report (ENIQ report No 41: “Probability of Detection Curves: Statistical Best-Practices”), (2) to justify the Rule-of-Thumb that a valid Probability of Detection (POD) *vs.* size curve requires a minimum of 60 targets for binary response (hit/miss) data, (3) to provide guidelines for the NDE practitioner in designing a study to assess the effectiveness of a binary response inspection system using POD *vs.* size curves.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LD-NA-25200-EN-N

Publications Office

ISBN 978-92-79-23018-9